Bayesian machine learning methods for predicting protein-peptide interactions and detecting mosaic structures in DNA sequences alignments

Wolfgang Lehrach

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2008

Abstract

Short well-defined domains known as peptide recognition modules (PRMs) regulate many important protein-protein interactions involved in the formation of macromolecular complexes and biochemical pathways. High-throughput experiments like yeast two-hybrid and phage display are expensive and intrinsically noisy, therefore it would be desirable to target informative interactions and pursue *in silico* approaches. We propose a probabilistic discriminative approach for predicting PRM-mediated protein-protein interactions from sequence data. The model suffered from over-fitting, so Laplacian regularisation was found to be important in achieving a reasonable generalisation performance. A hybrid approach yielded the best performance, where the binding site motifs were initialised with the predictions of a generative model. We also propose another discriminative model which can be applied to all sequences present in the organism at a significantly lower computational cost. This is due to its additional assumption that the underlying binding sites tend to be similar.

It is difficult to distinguish between the binding site motifs of the PRM due to the small number of instances of each binding site motif. However, closely related species are expected to share similar binding sites, which would be expected to be highly conserved. We investigated rate variation along DNA sequence alignments, modelling confounding effects such as recombination. Traditional approaches to phylogenetic inference assume that a single phylogenetic tree can represent the relationships and divergences between the taxa. However, taxa sequences exhibit varying levels of conservation, e.g. due to regulatory elements and active binding sites, and certain bacteria and viruses undergo interspecific recombination. We propose a phylogenetic factorial hidden Markov model to infer recombination and rate variation. We examined the performance of our model and inference scheme on various synthetic alignments, and compared it to state of the art breakpoint models. We investigated three DNA sequence alignments: one of maize actin genes, one bacterial (Neisseria), and the other of HIV-1. Inference is carried out in the Bayesian framework, using Reversible Jump Markov Chain Monte Carlo.

Acknowledgements

First, I would like to thank my principal supervisor, Dirk Husmeier, for the many enthused, enlightening, and more than occasionally adversarial discussions that were vital to my development as a researcher. In particular, his tolerance of me inevitably finding bugs close to paper deadlines was invaluable. I however suspect that any more announcements of "Well, I've got good news, and I've got bad news" will leave permanent mental scarring (future students take note). My thanks also go to my secondary supervisor, Chris Williams, who also provided assistance and ideas. My examiners Mark Girolami (external) and Amos Storkey (internal) took the time to (repeatedly) read, and understand my thesis, resulting in feedback that undoubtedly improved its comprehensibility and diplomacy.

The machine learning group at the Institute of Adaptive and Neural Computation has provided many stimulating and extensive discussions of the current work occurring in the field, and of my own work, instilling in me a deep interest in research. In particular, the members of the PIGs (Probability Inference Group) journal club introduced me to and performed useful critical analysis on many exciting developments. Biomathematics and Statistics Scotland, my other institute, also provided insights in more biological and classical statistics problems. A big thank you to my office mates for their companionship, the many resulting games of pool and pub trips (and the thus resulting knowledge of Scottish ales and whiskies), and the many interesting (occasionally even relevant!) office discussions. My flatmate, Timothy Hospedales, also deserves mention for his friendship and conversations over the course of the masters and the Ph.D. Also, of course, thanks must go to my family for their (lifelong) encouragement and understanding.

Last, but certainly not least, Charlotte Bowell, my girlfriend for her loving support and ability to keep me sane over the years of my Ph.D, and more.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein-Protein interactions

Proteins play a key role in almost all cellular processes (Gavin et al., 2006), and it is by their interactions that they carry out most of their key roles, such as building signalling networks, post translation modification of other proteins and forming structural components for the cell. Discovering the interactions between the proteins thus gives an insight into the function of the protein (Nabieva et al., 2005), and helps approaches such as systems biology to build quantitative models that can predict the behaviour of the cell (Stelzl and Wanker, 2006).

Recent experiments have attempted to characterise protein interaction networks and the complexes formed by groups of proteins on an organism-wide basis. Experimental methods used include yeast two-hybrid (Uetz et al., 2000; Ito et al., 2001; Giot et al., 2003), Tandem Affinity Purification (TAP, Krogan et al., 2006), and combinations of TAP and mass spectrometry (Gavin et al., 2006). See Shoemaker and Panchenko (2007) for a wider overview of experimental methods for detecting protein interactions, and of databases of known interactions. While high throughput methods can detect a substantial fraction of the interactions and complexes within an organism, they cannot explain the purpose of these interactions, nor how they are mediated. Low throughput methods like x-ray crystallography, electron microscopy, or electron tomography can determine the complex formed by proteins interacting (Russell et al., 2004). However, these methods have drawbacks. In order to perform x-ray crystallography, a sufficient quantity of the protein complexes is required, which then needs to be induced to crystallise. Electric microscopy and electric tomography on the other hand, suffer from low resolution, making interpretation of the protein complexes difficult.

Sometimes however, the underlying protein-protein interaction mechanism is simple enough that the cause of the interactions can be computationally inferred, as for instance when the interaction is mediated by a Peptide Recognition Module (PRM). PRMs are specialised compact protein domains that mediate many important protein-protein interactions. They are responsi-

ble for the assembly of critical macromolecular complexes and biochemical pathways (Pawson and Scott, 1997), and they have been implicated in carcinogenesis and various other human diseases (Sudol and Hunter, 2000). These domains bind to a specific short peptide sequence motif. In Chapter 3, we propose a novel discriminative method for discovering and distinguishing between the binding site motifs of these PRMs based on their interaction partners. Our method is applied to the SH3 domains in *Saccharmomyces cerevisiae*, where the interaction data comes from Tong et al. (2002).

The model that we proposed in Chapter 3 was only trained on peptide sequences that were found to interact with at least a single SH3 Domain, as its application is otherwise computationally impractical. In Chapter 4, we propose an additional, significantly less computationally costly method which can be efficiently applied to large numbers of sequences. This is due to its additional assumption that the underlying binding sites of different SH3 domains tends to be similar.

1.2 Phylogenetics and Comparative genomics

The binding site motifs of PRMs (e.g. SH3 domains) are short and degenerate. They are thus non-trivial to correctly detect, so additional clues to their location could be helpful. While mutations are a random process and thus occur in random positions, mutations in the functional regions of protein, like the binding site motifs of the PRM, tend to be more likely to be detrimental to the survival of the organism. Natural selection acts as a filter upon the mutations that will fixate in the population, so important regions will have less mutations and thus be more conserved. These conserved regions can be located by comparing the sequence to its homologues in other species. More conserved regions of the sequences are those where sequence is more similar to its homologues. See for instance Nimrod et al. (2005) who used this conservation (combined with the structure of the protein) to locate functionally important regions in proteins. To summarise: regions of proteins that are found to be more conserved should be made to have an increased *a priori* probability of containing the binding site motifs of the PRM. This can be incorporated using a suitable prior as proposed in Section 7.2.1.

As the concept of conserved regions is only defined in an evolutionary context, we need to understand the evolutionary history of a sequence. Understanding this history is the central problem of both phylogenetics and comparative genomics. Phylogenetics focuses on inferring the evolutionary relationships between a set of species, while comparative genomics models the rate variation that occurs along the sequence. Phylogenetics, apart from inferring the "tree of life", also has many useful applications in different fields such as epidemiology. For instance, Crandall (1995) showed how multiple HIV infections could be traced to back a single HIV positive dentist who had infected his patients. It was found that the HIV strains of the dentist

were closely related to those found in the patients, and not to other possible infection sources.

Modelling rate variation along sequences is central to the field of comparative genomics, where it can reveal functionally important regions. See for instance Margulies et al. (2007), who used a wide variety of experimental methods in an attempt to rigorously find every functional element along 1% of the human genome. They then compared these functional elements to the conserved areas found along these genome fragments, and found that most classes of functional sequence elements were enriched in conserved, as opposed to unconserved, regions. It was also found that significant numbers of each functional class occurred in non-conserved regions, and that 40% of the conserved regions did not overlap any known functional region.

Ideally, the evolutionary history of every nucleotide in every gene of every species could be traced. Not only would this be useful in identifying the binding sites of the PRMs, but it would also be highly revealing about the function of the nucleotides, as well as the selective pressure on the species that determined which mutations were kept.

Consider wanting to infer the evolutionary history of a set of DNA sequences, for instance the genomes of a set of species. Specific subsequences within the genomes can evolve in separate ways to the rest of the sequence due to recombination, gene duplication, retroviruses inserting themselves into the genome, etc. We will not focus on the evolution of species in this thesis, but only on modelling the evolution of the sequences. While we mention DNA sequences here, our description applies equally well to peptide sequences.

Given that each target sequence position has only one of the four possible nucleotides, it would naïvely appear that not enough information is preserved to be able to infer the evolutionary history. Furthermore, it is not known which positions in each of the sequences correspond to positions in the other sequences, because mutations can shorten or lengthen the sequences, and other processes such as recombination copy or move around large groups of nucleotides. These difficulties would appear to make the task impossible.

In practice, there is a strong *a priori* intuition that neighbouring positions in the sequences tend to share a similar evolutionary history. Given some suitable model, it should be possible to simultaneously infer the alignment of nucleotides (which nucleotides correspond between sequences) and their evolutionary history. However, simultaneously estimating both the alignment and the phylogenetic tree per alignment position is a difficult and unsolved computational problem. Most phylogenetic methods (including the method proposed in this thesis) instead start from an alignment of the nucleotides – see Lassmann and Sonnhammer (2005) for an example of a recent, high performance sequence alignment method.

Traditional methods in phylogenetics extend the assumption that neighbouring sequence positions have the same evolutionary history to its extreme, and assume that all nucleotides in a given sequence share the same evolutionary history. Recall that the binding site motifs would be expected to be more conserved (to have a lower average branch length in its phylo-

genetic tree). Without relaxing this assumption, this type of model would not help in detecting the binding sites motif, as all sites would show the same amount of evolutionary divergence. This assumption is also broken in certain bacteria and viruses as they undergo interspecific recombination. Here different strains exchange or transfer DNA subsequences, leading to a tree topology change, breaking the assumption that all positions can be modelled with a single phylogenetic tree. We propose a novel method for simultaneously characterising rate variation and recombination along alignments of DNA sequences.

1.3 Overview of the thesis

Chapter 2 reviews the current literature that describes phylogenetic methods for discovering sequence motifs and predicting protein-protein interactions. Chapter 3 introduces a novel discriminative method for discriminating between the binding sites motifs of PRMs. We investigate the performance of this method in distinguishing between the SH3 domain binding sites in *Saccharmomyces cerevisiae*. However, it is computationally impractical to apply this method to all sequences in yeast which do not bind to an SH3 domain. In Chapter 4 we cover an alternative, computationally efficient method for incorporating large numbers of sequences in finding these binding site motifs. Motifs tend to have a different rate of conservation from other parts of the sequence, suggesting that a phylogenetic approach might prove fruitful in helping to identify the motifs. We start by introducing some basic concepts in phylogenetics in Chapter 5. In Chapter 6 we propose a novel method for characterising rate variation and recombination along sequence alignments. In Chapter 7, we detail possible methods to combine these different sources of information, and conclude with a summary of the contributions of this thesis.

Part I

Protein-protein interactions

Chapter 2

Discovering motifs and predicting protein-protein interactions

2.1 Overview

In this chapter, we will review the current literature about discovering motifs and predicting protein-protein interactions. We start in Section 2.2.1 by reviewing methods for discovering motifs. Section 2.3 then shows how motifs and domains can be used to predict protein-protein interactions. Section 2.4 describes other *in silico* methods for predicting protein-protein interactions. Finally, Section 2.5 outlines how the methods described in this Chapter lead into our first discriminative model for simultaneously locating motifs and predicting protein-interactions.

2.2 Discovering motifs and locating domains

2.2.1 Proteins, Domains, DNA, and Motifs

In this thesis, the focus is on locating sequence motifs that occur on proteins. However, a lot of the methods for detecting motifs in DNA sequences are relevant. We start by introducing proteins and DNA sequences.

Proteins consist of one or more chains of peptides, where successive peptides in each chain are covalently bonded together. There are twenty possible peptides (also called amino acids) that can occur in each position in each chain. It is the arrangement of these peptides that gives the protein its shape, as different peptides interact with each other in different ways. The interactions, and thus the functions of a protein are determined by its shape which is ultimately determined by the sequence(s) of peptides. This description misses out various subtleties such as other molecules binding to the proteins and post-translational modifications of proteins – see for instance Alberts et al. (2001) for a detailed introductory text on molecular biology.

DNA sequences are long, stable chains of nucleotides bonded together that are generally used for long term storage of genetic information within an organism. These sequences code for everything within the cell (including proteins) and contain regulatory elements used to control the expression levels of the proteins. There are four different nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Again, see an introductory text on molecular biology such as Alberts et al. (2001).

A sequence motif generally refers to a very short (4-10 positions) conserved subsequence on either DNA or protein sequences. Sequence motifs are generally distinguishable by their biological significance or statistical over-representation. These sequence motifs occur for different reasons in protein versus DNA sequences, and we will quickly cover the most relevant occurrences of why DNA and protein (or peptide) sequence motifs occur.

Short peptide sequence motifs are commonly known as Linear Motifs. One particularly interesting example of such motifs are the binding sites of various domains called Peptide Recognition Modules (PRMs). These are known to be involved in forming many important complexes with organisms, and in various human diseases such as carcinogenesis. As PRMs are described in Section 3.3, we will not talk about them further here.

The most interesting occurrences of sequence motifs in DNA sequences has been traditionally considered to be in the upstream region of genes as these motifs can indicate the binding sites for Transcription Factors (TFs). However, recent research by Margulies et al. (2007) has shown that binding sites for TF can occur in the downstream region of a gene, or indeed almost anywhere on the genome. The binding of these TFs in turn activates or inhibits the transcription machinery, and hence the activity of that gene. This has been the source of significant research activity – Sandve and Drablos (2006) mention that over a hundred different methods for discovering DNA motifs in regulatory regions have been published in the last few years. We will not cover all of these methods, instead focusing on methods which are either illustrative of fundamental concepts in motif finding or the most applicable to discovering motifs in protein sequences.

Motifs can be defined for other characteristics than sequence: a structural motif is a small, repeatedly occurring substructure within the structures of proteins, where the amino acids involved do not have to be in consecutive sequence positions. Determining the structure of a protein is time consuming and expensive, and not all proteins have had their structure determined. We will focus on conserved elements which can be detected using only the sequences, namely domains and sequence motifs (motifs will now refer to sequence motifs from here on).

Domains are conserved substructures that repeatedly occur on different proteins, generally consisting of 40 to 350 successive amino acids (Alberts et al., 2001). We will not focus on domains as they tend to be easier to discover than peptide sequence motifs due to their extra length. Sequence motifs can be as short as only a few amino acids (or nucleotides) long

and thus can sometimes exhibit only a very weak statistical signal. Locating this weak signal amongst the noise can be a difficult computational problem. When looking for motifs with a weak signal, other repeatedly occurring subsequences that appear to be motifs might be detected in the noise – it is determining the significance and purpose of a potential motif that is the most difficult problem in motif discovery. The methods to detect and characterise domains also tend to be different as domains are more likely to be of variable length, and are thus often characterised by hidden Markov models. In this thesis we will focus on the harder case of discovering the short sequence motifs as opposed to discovering the longer domains. If the knowledge of which domains are present on a given protein is required, we will consult databases such as Interpro (Mulder et al., 2007). Interpro is a collation of many different databases of domains, as found by different domain discovering methods. Different scanning methods detect different subsets of the domains. Interpro attempts to characterise the underlying domains from the noisy and incomplete domain detections of the various domain discovery methods. If domain information is required for peptide sequences that do not have an entry in the database, tools such as InterproScan (Quevillon et al., 2005) search peptide sequences for occurrences of domains known within Interpro.

Methods for discovering the relatively short sequence motifs can be split into two main categories: methods that only look for statistically over-represented subsequences, and methods that explicitly incorporate additional information (e.g. gene expression data or protein-interaction data) to try and find motifs that are more likely to be biologically relevant. We will separately deal with DNA and peptide motif discovering methods that use additional information as the types of additional information useful for finding DNA sequence motifs differs from the types of additional information useful for finding protein motifs.

2.2.2 Discovering over-represented motifs

Most motifs are detected as overrepresented sequence fragments, which appear as outliers within some statistical model of the general characteristics of the sequence. This requires a model of the motif and at least an implicit model of the sequence. For instance, Pavesi et al. (2001) represented motifs as a consensus sequence, which is simply a list of successive nucleotides/peptides that represent the motif. Possible matches between a motif and a candidate subsequence were scored by the number of changes required for the motif and the candidate sequence to match. The advantage of such simple motif representations is that the search space is sufficiently restricted that all high scoring motifs can be efficiently found.

Not all motifs are well represented by a consensus sequence – in practice each motif position will have a different affinity for all possible nucleotides/peptides. This is ignored by the simpler consensus motif models, so we will instead focus on probabilistic models that describe the motif as a product of multinomial distributions, where each position in the motif is mod-

elled as an independent multinomial distribution. This model of a motif is called a Position Specific Scoring Matrix (PSSM), or Position Weight Matrix (PWM), and will be described in detail later.

PWM/PSSM motif models assume that the positions in a motif can be modelled independently. O'Flanagan et al. (2005) investigated how the assumption of independence between motif positions affects the performance of modelling regulatory motifs in DNA sequence. They used a computational approach to estimate protein-DNA binding energies by aligning the structure of the protein binding site with the different possible DNA binding motifs (see Section 2.4.1 for more detail about such methods). This allowed them to analyse the non-additive effects on the binding specificity of the protein. They discovered that such non-linear factors were caused by the folding of the target DNA sequence. Barash et al. (2003) suggests various more complex Bayesian mixture and tree based models of the motif. In a tree based model, the distribution over some sites in the motif are dependent on other sites in the motif. These models might be flexible enough to capture the variation caused by the non-additivity effects. However, it is not known to what extent such non-additivity effects occur with sequence motifs on proteins, and breaking the independence assumption makes the model significantly more difficult to fit and use. We will focus on methods that keep the assumption that the position of the motif can be modelled independently, but incorporating more complex motif models might be an interesting avenue of future research.

One of the best known methods for detecting over-represented sequence motifs in either nucleotide or peptide sequences is Multiple EM for Motif Elicitation (MEME; EM is Expectation Maximisation), described for instance in Bailey and Elkan (1995). MEME uses a product of multinomial distributions to describe the motifs, and a multinomial distribution to represent the background. The background refers to the background of the motifs, or all parts of the sequence that do not contain the motif. They define $\theta = [\theta_0\ \theta_1]$ to represent the parameters of their model, where θ_0 are the parameters of the background distribution and θ_1 are the parameters of the motif distribution. θ_0 and θ_1 are defined as follows:

$$\theta_0 = \begin{bmatrix} P_{a,0} \\ P_{b,0} \\ \vdots \\ P_{z,0} \end{bmatrix} \qquad \theta_1 = \begin{bmatrix} P_{a,1} & P_{a,2} & \cdots & P_{a,W} \\ P_{b,1} & P_{b,2} & \cdots & P_{b,W} \\ \vdots & \vdots & \ddots & \vdots \\ P_{z,1} & P_{z,2} & \cdots & P_{z,W} \end{bmatrix}. \tag{2.1}$$

$P_{a,0}$ to $P_{z,0}$ define a multinomial distribution that describes the background distribution, i.e. the distribution over all sequence positions that are not in a motif. The letters a to z are the set of all possible letters that can occur in the sequence. For instance, when modelling nucleotide sequences the letters would be: a =adenine, b=thymine, c =cytosine and d =guanine. Then, $P_{a,0}$ is the probability of the first letter occurring in the position not containing a motif (in the background), $P_{b,0}$ is the probability of the second letter, etc. Correspondingly, $P_{a,1}$ to $P_{z,1}$

define the probabilities of each letter occurring in the first motif position, $P_{a,2}$ to $P_{z,2}$ define which letters are expected in the second motif position, etc. This continues for all W positions in the motif. This representation of a motif is the aforementioned PSSM or PWM.

Bailey and Elkan (1995) introduced the set of latent (hidden) binary variables $Z_{i,j} \in \{0,1\}$, where $Z_{i,j} = 1$ if on the i^{th} sequence, the motif starts at the j^{th} position ($Z_{i,j} = 0$ otherwise) in order to model where the instances of the motif occur along the sequences. The log probability of a sequence X_i given that for one and only one j, $Z_{i,j} = 1$ was:

$$\log P(X_i|Z_{i,j}=1,\theta) = \overbrace{\sum_{k=0}^{j-1} \log P_{s_{i,k},0}}^{\text{Background}} + \overbrace{\sum_{k=0}^{W-1} \log P_{s_{i,j+k},k}}^{\text{Motif}} + \overbrace{\sum_{k=j+W}^{M} \log P_{s_{i,k},0}}^{\text{Background}}, \tag{2.2}$$

where $s_{i,k} = a$ shows that the a^{th} letter of the sequence alphabet occurs on the i^{th} sequence in the k^{th} position. This is a slightly different formulation from that used by Bailey and Elkan (1995), and is chosen to be more consistent with the notation used in Chapter 3.

Discovering motifs is equivalent to finding the optimal assignments to $Z_{i,j}$ and θ. This requires a model of how often the motif occurs in each input sequence. The authors introduced three such models: One Occurrence Per Sequences (OOPS), Zero or One Occurrence Per Sequence (ZOOPS), and Two Component Mixture (TCM). The OOPS and ZOOPS models are self-explanatory, while the TCM can represent zero or more motifs occurring along each of the sequences. The log likelihood of $X = \{X_1, \ldots, X_n\}$, a set of n sequences of length L[1] under the OOPS model was:

$$\log P(X,Z|\theta) = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j} \log P(X_i|Z_{i,j}=1,\theta) \right) + \frac{1}{n} \log m, \tag{2.3}$$

where $m = L - W + 1$ is the number of possible starting positions along the sequence for the motif and the $\log P(X_i|Z_{i,j}=1,\theta)$ term was defined in Equation (2.2). The likelihood of the sequences given $Z_{i,j}$ and θ under the ZOOPS model was:

$$\begin{aligned} \log P(X,Z|\theta) = &\left(\sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j} \log P(X_i|Z_{i,j}=1,\theta) \right) + \sum_{i=1}^{n} (1-Q_i) \log P(X_i|Q_i=0,\theta) \\ &+ \sum_{i=1}^{n} (1-Q_i) \log(1-\gamma) + \sum_{i=1}^{n} Q_i \log \frac{\gamma}{m}, \end{aligned} \tag{2.4}$$

where $Q_i = \sum_{j=1}^{m} Z_{i,j}$ indicates if the i^{th} sequence contains an instance of the motif, and γ is the probability of each input sequence containing the motif. The likelihood of the sequences under the TCM model was:

$$\begin{aligned} \log P(X,Z|\theta) = &\sum_{i=1}^{n} \sum_{j=1}^{m} (1-Z_{i,j}) \log P(X_{i,j}|\theta_0) + Z_{i,j} \log P(X_{i,j}|\theta_1) \\ &+ (1-Z_{i,j}) \log(1-\lambda) + (Z_{i,j}) \log \lambda, \end{aligned} \tag{2.5}$$

[1] For simplicity of exposition only – the model itself is not limited to equal length sequences.

where $X_{i,j} = \{s_{i,j}, s_{i,j+1}, \ldots, s_{i,j+W-1}\}$ and λ is the probability that any given position in the sequence contains the motifs. Under this model, every possible motif position is modelled almost independently[2] as either containing a motif or being a background position. The likelihood of a possible motif position given that there is no motif present is $P(X_{i,j}|\theta_0) = \sum_{k=0}^{W-1} \log P_{0,s_{i,j+k}}$, while the likelihood of a possible motif position given that there is a motif present is $P(X_{i,j}|\theta_1) = \sum_{k=0}^{W-1} \log P_{k,s_{i,j+k}}$.

The log likelihood described in either Equation (2.3), (2.4) or (2.5) was optimised by adjusting θ (where θ is augmented with γ or λ if appropriate) using the Expectation-Maximisation (EM) algorithm.

First, $\theta^{(0)}$, a starting value for θ is chosen. Then $P\left(Z|X,\theta^{(0)}\right)$, the posterior distribution of Z, is computed. Given that in this case this posterior distribution is both conjugate and in the exponential family, it can be described by a small set of sufficient statistics. For the motif finding problem, the posterior distribution corresponds to the probability that each position in each sequence contains the motif. Computing this posterior distribution is called the E-step. These distributions over Z are then used to find the optimal values of θ that maximise the log likelihood in the M-step:

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_{\theta} \langle \log P(Z,X|\theta) \rangle_{Z|X,\theta^{(t)}} \\ &= \arg\max_{\theta} \int_Z \log P(Z,X|\theta) P\left(Z|X,\theta^{(t)}\right) dZ, \end{aligned} \tag{2.6}$$

where $\langle f(X) \rangle_X = \int_X f(X) P(X) dX$ corresponds to taking the expectation of a function of a variable with respect to a distribution. $\theta^{(t+1)}$ is in turn used to calculate the posterior distribution $P\left(Z|X,\theta^{(t+1)}\right)$, which is in turn used to find $\theta^{(t+2)}$. This is repeated until the difference $\left|\theta^{(t+1)} - \theta^{(t)}\right|$ falls below some threshold, which is taken as an indication that the algorithm has converged. EM is guaranteed to converge to some local optimum or saddle point of the maximum likelihood function, as shown by Dempster et al. (1977). In practice, many different initialisations from different possible motif starting positions are tried and run for a small number of steps. Whichever initialisation shows the most promise is then run until convergence. Given a motif has been discovered, all parts of the sequence containing this motif are removed from the sequences, and the algorithm is rerun to locate the next motif.

The quality of the motifs found is still highly dependent on the initialisation strategy used due to the greedy nature of the optimisation. An alternative method is to use Gibbs sampling for motif finding as described for instance by Lawrence et al. (1993). Here, θ is sampled given Z, then Z is sampled given θ, then θ is sampled given Z, etc. This will eventually yield samples for Z and θ from posterior distributions of Z and θ and is thus in theory independent of the starting initialisations of Z and θ. In practice, Gibbs sampling can require large numbers of

[2]Possible motif position are not fully independent as extra care is taken to ensure that motif instances do not overlap.

iterations, and thus computational time, to switch between different motifs. Gibbs sampling can thus in practice still be dependent on the initialisation.

The simplistic model for the background (any sequence position that does not contain a motif) used by Lawrence et al. (1993) and Bailey and Elkan (1995) is too naïve to fully capture the behaviour of the sequences. Thijs et al. (2001) implemented a third order Markov model of the background sequence positions, where the distribution of the current site depends on the last two sites. The authors showed that their more realistic background model improved the ability of the model to distinguish between DNA regulatory motifs and background sequences compared to the more simplistic background model.

Leung and Chin (2005) introduced a method that is guaranteed to find the best possible PSSM motif up to some given error δ. The lower the required error δ, the greater the amount of computational time required. Their method is designed to be applied to DNA sequences, and so will be impractical to apply to protein sequences as the search space will be too large.

Down and Hubbard (2005) introduced a novel method to locate motifs using nested sampling (see Skilling, 2006 for an overview of using nested sampling for Bayesian inference). Nested sampling maintains an ensemble of solutions and so is a different type of approach than the Expectation Maximisation and Gibbs methods. At each iteration, the solution with the lowest likelihood is discarded. A new solution is sampled from the prior, under the constraint that its likelihood must be equal to or greater than that of the discarded solution. Down and Hubbard (2005) sampled from this constrained distribution by duplicating an existing solution, then using standard Metropolis-Hastings to update the new duplicate to a new position within the constrained prior.

Down and Hubbard (2005) mention that nested sampling is similar to simulated annealing, an alternative optimisation strategy where updates are randomly proposed. In simulated annealing, the probability of accepting a worse solution depends on the current temperature and how much worse the proposed solution is. The higher the temperature, the more likely it is that worse solutions are accepted. Initially the temperature is high and then decreases during the run of the simulation until no more solutions that are worse are accepted. Given this decrease is gradual enough, simulated annealing is guaranteed to find the optimal solution. Nested sampling has the advantage of not requiring such a temperture shedule, and Down and Hubbard (2005) claimed that nested sampling is liable to find the global optimum in a single run, given a large enough ensemble. However, this assumes that a suitable method exists for sampling from this likelihood-constrained prior. In practice, this can be very difficult, and is still an open question in sampling methodology (Girolami, 2007).

In most motif discovery methods, multiple motifs are often found by first finding a single motif, removing the subsequences containing that motif, then finding another motif, removing those subsequences, etc. This is done for instance in MEME. In contrast, the method of Down

and Hubbard (2005) simultaneously models all motifs that occur within the sequences by modelling the sequences as a Hidden Markov Model (HMM) where the hidden state represents the background state, or a motif position. The model can thus transition into any known motif repeatedly along each of the sequences. See also Down et al. (2007) for a large scale application of this method to discovering novel regulatory motifs in *Drosophila melanogaster*.

There are other metrics for scoring motifs than the log likelihoods shown in Equation (2.3), (2.4) and (2.5). For instance, Ng et al. (2006) focused on finding faint motifs that are difficult to distinguish from randomly generated background sequences. They optimised the Incomplete Likelihood Ratio which is the probability of an OOPS model (as in Equation (2.3)) as opposed to all the sequences being generated from the background distribution. They claim that this increases the ability of the method to detect faint motifs. However, they only tested the motif finder on randomly generated background sequences – real biological sequences might also have more underlying structure, which might confuse this scoring measure.

Most of the motif finding algorithms outlined above assume a target set of sequences has been selected, where this target set of sequences are suspected to contain relevant motifs. However, it is not always possible to directly split a set of sequences into target and background sets in such a fashion. Instead, one might have a ranking as generated from a biological measurement such as ChIP-chip data (Chromatin ImmunoPrecipitation on a chip – see for instance Ren et al., 2000), where motifs that are relatively more abundant at the top of the ranking are likely to be interesting. Eden et al. (2007) introduced a method for finding motifs that are over-represented at the top or bottom of such a ranking.

A thorough investigation into discovering regulatory motifs from DNA sequences through the motif over-representation was carried out by Tompa et al. (2005). The authors quantitatively compared thirteen DNA sequence motif finding algorithms that look for over-represented short sequences that occur in the upstream region of genes, as these are likely to be good candidates for regulatory elements. They found that there was some complementarity between the set of motifs found with the various methods.

2.2.3 Discovering regulatory motifs by incorporating other information sources

Unlike the methods described in Section 2.2.2, the motif discovering methods described in this section explicitly use some sort of additional information to help locate regulatory DNA motifs.

Segal et al. (2002) proposed a probabilistic model that models experimentally observed gene expressions as the result of the presence of regulatory motifs in their upstream regions. Of particular interest is their discriminative model for discovering motifs. Instead of modelling the motif and the background distribution as shown in Equation (2.1), they directly modelled the likelihood ratio of each amino acid appearing in a given motif position as opposed to anywhere else in the motif or the background – see Section 3.4.2.

Transcription factors do not necessarily bind only to a single motif. Instead, multiple motifs can bind to the same transcription machinery. It is this binding to the same machinery that forces the motifs to be in spatial proximity to each other. Such groups of motifs are called cis-Regulatory Modules (CRMs), and the knowledge that multiple motifs occur in spatial proximity can aid in detecting them. Segal and Sharan (2004) proposed a novel discriminative model that exclusively focuses on pairs of motifs that are spatially correlated while ignoring other non-correlated motifs.

See Sandve and Drablos (2006) for a large overview of regulatory motif discovery methods in DNA sequences, organised by the type of contextual information used.

2.2.4 Discovering peptide sequence motifs using additional information

Short peptide sequence motifs are commonly known as Linear Motifs. These include the binding sites for Peptide Recognition Modules (PRMs), which are domains that bind to short peptide sequence motifs, and are described in detail in Section 3.3.

Reiss and Schwikowski (2004) proposed a model to help disambiguate the binding sites of a set of closely related PRM domains. This is a difficult problem as the motifs representing the various binding sites of the PRM domains are closely related, and so distinguishing between these binding sites is difficult. Searching for a motif amongst the sequences that bind to each individual SH3 domain yields too few examples to be able to find the binding site. Correspondingly, searching for motifs that occur in the sequences that bind to any SH3 domain merges together similar binding sites. By explicitly taking into account the underlying similarity between the binding sites of the different SH3 domains, the authors managed to more accurately model the binding site motifs of the SH3 domain. Their model is covered in detail in Section 3.4.1, and will not be described further here.

Neduva et al. (2005) carried out large-scale scans for linear motifs using whole proteome protein interaction datasets. They systematically scanned for over-represented linear motifs amongst the interaction partners of each protein. As this does not always yield enough results to determine the relevance of the motifs found, they also use as an alternative target set all proteins that were found to interact with any protein that contained a particular domain. They referred to this as a domain interaction set. Various motifs were found only in either the protein interaction set or domain interaction set, demonstrating the importance of examining both datasets.

Neduva et al. (2005) designed a simple scoring function that looked for significantly over-represented motifs in the target set as opposed to the background set. This scoring function was based on the binomial distribution. This is not the main novelty of their method, which instead comes from using protein-protein interaction datasets to select interesting target sets for their motif discovery algorithm. Their method is available on-line at `http://dilimot.embl.de/` (see Neduva and Russell, 2006 for more details of the web service).

In the model of Neduva et al. (2005), all positions in the motif are represented as either an amino acid or a wild card. Hence, any position in the motif which binds to a subset of amino acids has to be represented as a wild-card, leading to an over-estimation of the number of occurrences of the motifs in the background set, and reducing the statistical significance of the motifs found. Hence, a more detailed model of the motifs may increase the sensitivity of the method at discovering such motifs.

2.3 Predicting protein interactions by looking at domain interactions

In this section, we will review methods for explaining and thus predicting protein-protein interactions as the result of interactions between the domains present on the proteins. Within this section, the term domain will refer to any domain, motif, or otherwise identifiable sequence-signature on a protein. This does not include simple secondary structure elements such as α-helices and β-sheets as they occur on almost all proteins, and thus their presence or absence is not very informative for predicting interactions.

Methods for predicting protein-protein interactions from domains generally attempt to explain the observed interactions in terms of underlying domain-domain interactions. They can be broadly classified into three separate classes: heuristic methods that look for domain-pairs that frequently occur in protein-interaction pairs, probabilistic methods that infer which domain interactions best explain the observed protein interactions and "black box" methods that focus on maximum predictive performance. We will now cover each type of method in turn.

2.3.1 Domain interactions are independent

An early model for explaining protein interactions in terms of domain-domain interactions was introduced in Sprinzak and Marglit (2001). The authors attempted to quantify the amount of information about domain-domain interactions that can be extracted from protein-protein interaction pairs, and to build a predictor of protein-protein interactions from the discovered domain interactions. Let I_S be the set of interacting proteins pairs, where $(s_a, s_b) \in I_S$ implies that proteins s_a and s_b are interacting. The proteins were represented as the set of domains that they contain, so duplicate copies of the same domain were ignored. The bag of domains for each protein was extracted from Interpro (Mulder et al., 2003). Any proteins which were not found to contain any known domains were excluded from the analysis. Sprinzak and Marglit (2001) then counted the potential number of protein-protein interactions that each pairing of domains could be causing:

$$M_{i,j} = \mathbb{I}(i \leq j) \sum_{(s_a,s_b)\in I_S} \mathbb{I}((D_i \in s_a \wedge D_j \in s_b) \vee (D_j \in s_a \wedge D_i \in s_b)) \qquad (2.7)$$

where $\wedge$ is a logical-and, $\vee$ is a logical-or, and $\mathbb{I}(\cdot)$ is the indicator function. $\mathbb{I}(\cdot) = 1$ if the condition inside the brackets is true, else it evaluates to 0. Only upper diagonal matrix elements are filled in. Given the scoring matrix $M_{i,j}$, they looked for over-represented domain pairings by searching for domain-pairings with a high log-odds score. A high log-odds score between a pair of domains was then taken to indicate that they are likely to interact. The log-odds score is defined as:

$$L_{i,j} = \log_2\left(\frac{P_{i,j}}{P_i P_j}\right), \tag{2.8}$$

where:

$$P_{i,j} = \frac{M_{i,j}}{\sum_i \sum_j M_{i,j}} \qquad P_i = \frac{\sum_j M_{i,j}}{\sum_i \sum_j M_{i,j}} \qquad P_j = \frac{\sum_i M_{i,j}}{\sum_i \sum_j M_{i,j}}. \tag{2.9}$$

A protein pair (s_a, s_b) is predicted to interact if there exist domains D_i and D_j such that $((D_i \in s_a \wedge D_j \in s_b) \vee (D_j \in s_a \wedge D_i \in s_b)) \wedge L_{i,j} > 2$. In words, proteins s_a and s_b interact if there is a pair of domains D_i and D_j – with one domain on s_a and the other domain on s_b – that have at least a log-odds score of 2 to be interacting. This heuristic predicts a small number of proteins interaction pairs, with a relatively high proportion of accurately predicted interactions. However, only a small number of interactions were predicted between the proteins, and the authors did not estimate the proportion of protein interactions that are missed by their method.

The authors claimed that this log-odds score shown in Equation (2.8) measures the enrichment of domain-pairings occurring in protein-interactions, as compared to the number of the domain pairings expected to occur at random, and that it is hence a good indicator of domains that are likely to be interacting. However, random pairings between domains were only considered between domains found on proteins that were involved in interactions. All instances of domain-pairings between non-interacting proteins are ignored, which may significantly alter which domains would be expected to interact. Thus, their claim that the log odds score from Equation (2.8) is a good predictor of domain interactions is arguably suspect, as ignoring these domains between non-interacting proteins would give occasionally give inflated significance values to the predicted domain pairings.

Most statistical tests such as the χ^2 squared test are not immediately applicable, as they would test if the variables P_i and P_j are independent. This would involve summing over all possible settings of these variables, while here we are interested in how likely it is that a pair of domains (i.e. a specific settings of the variables) is enriched compared to what would be expected from how often the individual domains occur, where this enrichment is taken as a sign that the domains are interacting. If the problem is restated such that predicting an interaction between each possible combination of domains involved an independence test between two variables, then such tests would be applicable and thus could be used to provide confidence values for each of the predicted domain interactions. Alternatively, their method could be improved by using an alternative to the heuristic minimum log-odds score of 2. This threshold

could, for instance, be set with cross-validation to get the optimal classification performance or the lowest rate of false positives. However, due to the reasons outlined in the last paragraph, this would be of limited utility as it would still ignore information from non-interacting protein pairs.

Deng et al. (2002) introduced a model for inferring domain-domain interactions. They claimed that Sprinzak and Marglit (2001) introduced the following association measure:

$$A(D_i, D_j) = \frac{I_{i,j}}{N_{i,j}}, \tag{2.10}$$

where $I_{i,j}$ is the number of interacting protein pairs where one protein contains domain D_i and the other protein contains domain D_j. $N_{i,j}$ is the total number of interacting/non-interacting pairs of proteins where again one protein contains domain D_i and the other protein contains domain D_j. The measure in Equation (2.10) is fundamentally different from the actual measure introduced in Sprinzak and Marglit (2001), as can be seen by noticing that the measure in Equation (2.10) takes into account all possible protein interactions pairs which where not found to interact. In contrast, the measure of Sprinzak and Marglit (2001) shown in Equation (2.8) ranges over all domain pairings resulting from proteins involved in interactions only. This can be seen by referring back to the definition in Equation (2.7).

The apparently incorrect measure from Equation (2.10) which was introduced by Deng et al. (2002) has been re-used by other papers, such as for instance Hayashida et al. (2003), Hayashida and Ueda (2004), Guimaraes et al. (2006), and Huang et al. (2007). All these papers claimed that this is the measure introduced in Sprinzak and Marglit (2001). Additionally, Gomez et al. (2003) appears to have used yet another mutually exclusive definition of the association method of Sprinzak and Marglit (2001).

Gomez et al. (2003) introduced an attraction-repulsion model in which the domains can repulse as well as attract each other. The probability of an interaction between the domains D_i and D_j under the *attraction-repulsion* model was defined as:

$$P(I_{i,j} = 1) = \frac{n^+_{i,j} + \Psi/2}{n^+_{i,j} + \gamma n^-_{i,j} + \Psi}, \tag{2.11}$$

where $n^+_{i,j}$ is the number of times domains D_i and D_j occurred in an interacting protein pair and $n^-_{i,j}$ is the number of times domains D_i and D_j occurred in a non-interacting protein pairing. $I_{i,j} \in \{0,1\}$ is a binary variable where $I_{i,j} = 1$ indicates that domains i and j interact. γ is a normalisation constant that was set such that the average probability of two domains interacting was 0.5, and Ψ is a pseudo-count to ensure that the count is non-zero. The probability of an interaction between proteins a and b was defined to be:

$$P(\varepsilon_{a,b} = 1) = \arg \max_{p(D_i,D_j)} \left| p(D_i, D_j) - 0.5 \right|, \tag{2.12}$$

where D_i ranges over domains in protein a, D_j ranges over domains in protein b, and $\varepsilon_{a,b} \in \{0,1\}$ is a binary variable where $\varepsilon_{a,b} = 1$ indicates that proteins a and b interact

Gomez et al. (2003) found that their method outperformed applying a Support Vector Machine (SVM – see Section 2.3.3) to the set of domains on each protein and that applying an SVM in turn outperformed the association method of Sprinzak and Marglit (2001). Gomez et al. (2003) also claimed that their heuristic method took significantly less training time and memory compared to applying an SVM. In order to over-come the limitations of representing proteins as a set of known domains, Gomez et al. (2003) investigated augmenting the information about which domains are present with how many times each of the possible 4-tuples of successive amino acids occurred. In order to reduce the feature space, the amino acids were grouped together according to their properties. Adding this extra method of characterising the sequence slightly improved the performance of their model compared to only taking into account the known domains. A possible reason for this is that such short features occur so frequently that they would be unable to describe any longer sequence elements like a novel domain, and thus that the knowledge of the biophysical properties of the amino acids introduced by the grouping is lost.

While Gomez et al. (2003) mentioned the latent variable method of Deng et al. (2002), they fail to compare the performance of their *attraction-repulsion* method to the latent variable model of Deng et al. (2002). Hence, it is not possible to judge if their heuristic scheme outperforms Deng et al. (2002). For instance, the latent variable model of Deng et al. (2002) may outperform the attraction-repulsion model of Gomez et al. (2003) as the attraction-repulsion model does not into account that some of the putative interactions can be explained away – see the example at the top of Section 2.3.2 for a detailed example. As the tuple representation appeared to contribute only slightly to the performance of the model, it would be unlikely to affect if Gomez et al. (2003) outperforms Deng et al. (2002).

Additionally, Gomez et al. (2003) appear to have used the slightly incorrect interpretation by Deng et al. (2002) of the association method of Sprinzak and Marglit (2001). This may have affected their comparison with the association method, but they would be expected to outperform the association method as Gomez et al. (2003) explicitly take into account domain pairings between non-interacting proteins, as well as allowing domains to be repulsed by each other.

Gomez and Rzhetsky (2002) incorporated global network degree priors into the inference of protein-protein interactions. The number of interactions that each protein is involved in can be approximated by a scale-free distribution, which the authors incorporated into their inference procedure. This required the use of Reversible Jump (RJ) Markov Chain Monte Carlo (MCMC) methods to predict which proteins interact, as each pair of protein-protein interactions were no longer independent. The probability of two domains interacting was simply estimated from

how often the given combination of domains occurs in interacting protein pairs.

Ng et al. (2003) proposed a model that brings in multiple data-sources like protein complex information and domain fusions to determine which domains interact. They use a probabilistic confidence scheme to integrate these disparate confidence sources.

2.3.2 Inferring the underlying domain-domain interactions

The methods outlined in the last section treated domain-domain interactions locally, i.e. they ignore that interactions can be explained away by other domain-domain pairings than the single domain-domain pairing being looked at. For an illustration of why this is a problem, consider a set of proteins $s_a, \ldots, s_z$ which contain the following domains: $s_a = \{D_i\}$, $s_b = \{D_j, D_k\}$ and $s_c = s_d = \cdots = s_z = \{D_k\}$. Consider the set of interacting proteins pairs: $\{(s_a, s_b), (s_a, s_c), (s_a, s_d), \ldots (s_a, s_z)\}$. We wish to infer how likely it is that each pair of domains interact.

The simple heuristic methods explained in Section 2.3.1 would predict that domains D_i and D_j interact, and that domains D_i and D_k interact. However, if D_i and D_k interact, then we have already explained the interaction between s_a and s_b. Hence, a proper probabilistic method should be unsure if D_i and D_j interact, as s_a and s_b would interact in any case due to the pairing of D_i and D_k.

Deng et al. (2002) formulated a probabilistic model where the probability of each possible pairing of domains interacting were modelled as hidden variables. The visible (or known) variables showed if each of the possible pairings of the proteins were found to interact within the yeast two-hybrid experiment. The task is to infer the hidden variables (domain-domain interactions) given visible variables (protein-protein interactions).

The central assumptions of Deng et al. (2002) were that all domain-domain interactions are independent of each other and that all protein interactions are caused by a pair of domains interacting. Under these assumptions, they modelled the probability of proteins i and j interacting as:

$$P(P_{i,j} = 1) = 1 - \prod_{D_{m,n} \in P_{i,j}} (1 - \lambda_{m,n}), \tag{2.13}$$

where $\lambda_{m,n}$ is the (to be inferred) probability that domains m and n interact, and $P_{i,j} \in \{0,1\}$ is a binary variable where $P_{i,j} = 1$ indicates that proteins i and j interact.

One of the problems with the association method of Sprinzak and Marglit (2001) is that the noise inherent in the experiment techniques is not modelled. Deng et al. (2002) tackled this by modelling the effect of the experiment noise on the observed protein interactions. They defined $O_{i,j} \in \{0,1\}$ and $P_{i,j} \in \{0,1\}$ as binary variables, where $O_{i,j} = 1$ indicates that an interaction between proteins i and j was observed in the experiment, while $P_{i,j} = 1$ indicates that an interaction actually occurs *in vivo*. They then defined the rate of false positives caused by experimental

error as $fp = P(O_{ij} = 1|P_{ij} = 0)$ and the rate of false negatives as $fn = P(O_{ij} = 0|P_{ij} = 1)$. The map between observed and true interactions was thus defined to be:

$$P(O_{ij} = 1) = P(P_{ij} = 1)(1 - fn) + (1 - P(P_{ij} = 1))fp, \tag{2.14}$$

where fn and fp were estimated from the difference in the number of observed and expected protein interactions and the error rates of yeast two-hybrid (the experimental method that was used to generated the dataset). The probability of the observed interactions, also called the likelihood, was thus:

$$L = \prod_{i,j} (P(O_{i,j} = 1))^{O_{i,j}} (1 - P(O_{i,j} = 1))^{1-O_{i,j}}, \tag{2.15}$$

where these terms are defined in Equation (2.13) and (2.14).

Deng et al. (2002) used the Expectation Maximisation (EM) algorithm to infer which underlying domain-domain interactions (i.e., the values of $\lambda_{i,j}$) would explain the observed protein-protein interactions. Once inferred, these values of $\lambda_{i,j}$ were used to predict interactions for novel pairings of proteins using Equation (2.13), where novel combinations of domains were not predicted to interact. Deng et al. (2002) showed that their method out-performs the association method of Sprinzak and Marglit (2001) – however, this comparison is arguably suspect, as Deng et al. (2002) appear to have used an incorrect definition of the association measure, as mentioned before. However, this would be unlikely to invalidate their conclusions due to the association method of Sprinzak and Marglit (2001), as their interpretation actually incorporates information from non-binding interaction pairs, which could well improve the performance of the associsation method.

The likelihood based approach used by Deng et al. (2002) did not express the uncertainty over how likely it is that a pair of domains interact. Consider again the example at the beginning of this section – their model cannot express that it is impossible to determine if domains D_i and D_j interact. In contrast, a fully Bayesian inference scheme (see for instance Bishop, 2006) captures such uncertainty by calculating posterior distributions for the variables. These posterior distributions help to identify good candidate domain-domain interactions, as we can examine the confidence of the $\lambda_{i,j}$ predictions, and select those $\lambda_{i,j}$ which are confidently predicted to be close to 1. Taking into account the uncertainty over the $\lambda_{i,j}$ variables should make the predicted protein interactions more dependent on the confidently predicted domain-domain interactions. This more rigorous approach should increase the performance of the model. One of the central problems in Bayesian statistics and Machine Learning is to build methods that can efficiently infer posterior distributions and marginal likelihoods that are of interest in a model (see for instance Bishop, 2006 or MacKay, 1992).

Alternative, non-Bayesian methods, could also be used to express uncertainty over the predictions. For instance, bootstrapping (Hesterberg et al., 2005) could be used to estimate

these confidence intervals. This involves repeated sampling from the dataset with replacement, and observing the effect on the estimated values of interest. However, these methods also lose other benefits of the Bayesian paradigm, such as the inherent regularisation of Bayesian models and Bayesian model selection and interpolation (again, see MacKay, 1992).

Yeast two-hybrid was used to generate the protein-protein interaction datasets for all the methods described – see for instance Twyman (2004) for a description of this method. Yeast two-hybrid is an experimental high-throughput method which is known to be noisy (Ito et al., 2001). This noise can be reduced by having many datasets that cover the same interactions. Then, the frequency of an interaction being found in the datasets indicates a relative confidence of each interaction occurring. This was not taken into account by Deng et al. (2002), as they only incorporated the overall uncertainty of the experimental method, not the uncertainty of each individual interaction occurring.

Hayashida et al. (2003) introduced an alternative, non-probabilistic model which takes the relative uncertainty of each interaction into account, slightly out-performing Deng et al. (2002) when this extra data is available. However, Deng et al. (2002) could easily be modified to incorporate this information by adjusting fn and fp on a per possible interaction basis, and a more rigorous probabilistic approach should ultimately exhibit a higher performance than the heuristic approach of Hayashida et al. (2003).

Huang et al. (2007) introduced an alternative paradigm for discovering domain-domain interactions by reducing the problem of predicting protein-protein interactions to discovering a weighted set of domain-domain interactions. These domain-domain interactions have to cover the protein-protein interactions with the maximum specificity. While they did not appear to perform significantly better than the latent variable model of Deng et al. (2002), they claimed that their method is faster.

Expanding on the model of Deng et al. (2002), Wang et al. (2004) explicitly took into account that domains can be buried or inactive in the proteins, and that protein interactions can occur for other reasons than interactions between the known domains. Their interpretations of a domain being inactive include it not being on the surface of the protein and hence not part of the binding site. An alternative interpretation is that instance of the domain is too evolutionarily distant from the other domains it was grouped with. While this made the model more powerful, it also made even the EM algorithm intractable. The authors coped with this by introducing a branch and bound method to approximate the E-step.

In order to check their predictions of which domains are active versus buried, the authors looked at proteins co-crystallised in the PDB database of protein structure (H.M.Berman et al., 2000). Domains on proteins involved in these interactions were determined to be active if any of their peptides where within a small distance of the interaction partner. They found that domains which they predicted to be active were more likely to be close to the binding partner.

It is not possible to say if the performance at predicting protein interactions has significantly improved as they do not attempt to use their model for this task.

2.3.3 Black box predictors

We will review methods in this section that use black box classifiers such as Support Vector Machine (SVMs) or neural networks to predict the protein-protein interactions. One of the principal advantages of such methods is that is they are not limited to modelling the protein interactions as the result of two domains interacting and that more information about each protein can easily be taken into account. However, this flexibility comes at the cost of interpretability, and it is almost impossible to understand by which criteria these models perform their predictions – hence the name "black box" classifiers.

Training an SVM is computationally efficient, and so is using the resulting trained model to perform classification. However, one of the disadvantages of SVMs is that they are not inherently probabilistic, and hence do not produce proper uncertainty intervals over their predictions. These probabilities be obtained in a post-hoc manner as done for instance by Platt (2000). Gaussian Process are an alternative, more rigorous, probabilistic kernel based method – see for instance Rasmussen and Williams (2005).

One of the earlier computational methods for predicting protein-protein interactions from protein sequences is the SVM method of Bock and Gough (2001). Unlike the other methods described, their method doesn't use information about the presence of domains. Instead, it predicts interactions directly from the sequences of the proteins involved. Bock and Gough (2001) defined a feature vector which represented a pair of interacting proteins. Each individual protein was represented as a vector of its amino acids, and each amino acid was in turn represented as a vector of its biophysical properties. Lots of biophysical properties of amino acids are known, such as the hydrophobicity, charge and surface tension (May, 1999).

Bock and Gough (2001) defined $\mathbf{v}_i^A$ to be the vector of the biophysical properties of the i^{th} amino acid along protein A. These vectors cannot be concatenated together as different proteins are of different lengths, so Bock and Gough (2001) mapped all proteins to a feature space of length N using nearest neighbour interpolation. This was done as follows: a protein of length M was represented as $\varphi_A = \mathbf{v}^A_{f(1)} \oplus \mathbf{v}^A_{f(2)} \oplus \cdots \oplus \mathbf{v}^A_{f(N)}$, where $\oplus$ is used to denote concatenation of vectors. The map $f(i) = \left[(i-1)\frac{M}{N} + 1\right]$ went from $i \in \{1,2,\ldots,N\}$, a position in the feature space to a position along the protein. $[\cdot]$ is used to represent integer rounding in the mapping.

A particular pair of interacting proteins A and B was represented as $\varphi_A \oplus \varphi_B$. An equal number of negative protein-protein interactions were generated by shuffling the protein sequences to destroy patterns indicative of interactions. These positive and negative example were then presented to the classifier. After the classifier had been trained on all positive and negative examples, it could then predict if an unseen pairing of proteins would interact.

The method of Bock and Gough (2001) is intuitively unsatisfying, amongst other reasons due to the rescaling of the proteins required. Any informative region would change in its representation due to unrelated extra folds occurring on the protein.

Dohkan et al. (2004) proposed an improved SVM based method by re-introducing domain information to the prediction of protein interactions. This was done by encoding each protein as $\hat{\varphi}_A = \varphi_A \oplus d_1^A \oplus d_2^A \oplus \ldots \oplus d_n^A$, where d_i^A is the number of times that the i^{th} domain occurs on protein A. An interacting pair of proteins was then represented as $\hat{\varphi}_A \oplus \hat{\varphi}_B$ and $\hat{\varphi}_B \oplus \hat{\varphi}_A$. This is unlike the method of Bock and Gough (2001) who apparently only used a single ordering $\psi_A \oplus \psi_B$. Instead of the peptide shuffling method used by Bock and Gough (2001), Dohkan et al. (2004) used all pairing of proteins which were not found to interact as the set of negative examples.

Dohkan et al. (2004) also investigated the effect of presenting various other features of the protein to the classifier in the same manner by augmenting φ with other protein features such as localisation data, amino acid composition, etc. They found that this extra information increased the performance of the methods, and that their method out-performed the method of Deng et al. (2002).

Ben-Hur and Noble (2005) extended this type of SVM approach by combining many different kernels that describe how similar two pairs of proteins are. They examined an extensive range of kernels that described the similarity between proteins. This included kernels based on non-sequence information including Gene Ontology (GO) annotations, the presence of interactions between paralogs within other species and mutual clustering coefficient, which measured how distant the proteins were within the interaction graph. The authors found that combining many different measures improved the overall performance of the classifier. The authors also incorporated different interactions datasets (high throughput methods generally have a lower reliability) by allowing a lower penalty for mis-predicting a low reliability interaction than a high reliability interaction.

Ben-Hur and Noble (2005) also argued that while the performance of non-sequence kernels at predicting protein-interactions was not much worse than the sequence based kernels, the non-sequence based kernels were not good at distinguishing between real interactions and co-complexed proteins. This is in contrast to the sequence based features which depend more on the properties of the interaction site itself. Overall, Ben-Hur and Noble (2005) exhibit promising levels of performance in predicting protein-protein interactions.

2.4 Other methods for predicting interactions

In this section, we will briefly review other methods for predicting protein-protein interactions. Within the scope of this thesis, these methods could prove useful for reducing the noise in the

interaction datasets. For instance, (Uetz et al., 2000) used yeast-two hybrid to map all protein interactions that occur in yeast. When Ito et al., 2001 repeated these experiments, they found only a small (10%) overlap of their results with those of (Uetz et al., 2000). Due to the large number of potential interactions (any protein could interact with any other protein), even a very low rate of false positives can swamp the true interactions. Hence, independent predictions of interactions could help to corroborate which interactions actually occur, and provide suitable prior information about potential protein-protein interactions.

2.4.1 Structural methods

The methods that have been discussed until now have not involved looking at the three dimensional structures of the protein. These structures are recorded in databases such as the PDB (see H.M.Berman et al., 2000). The protein structures can be used to predict interactions by attempting to optimally align the structures of these proteins that are suspected to interact. If the interaction surfaces between the structures align, then it may possible to predict that an interaction occurs between these proteins. In practice, this can be significantly more difficult than finding the best alignment between two rigid bodies as proteins can flex when they interact.

A yearly competition called Critical Assessment of PRediction of Interactions (CAPRI – see Janin, 2002 or the website at `http://capri.ebi.ac.uk/`) is held to assess the state of the art in the protein docking methods, and their ability to predict interactions. We will not cover these methods in detail here, as the focus of this thesis is on directly using sequence information. As an example, Davis et al. (2006) use a protein-protein docking method to predict protein complex formation.

It should be noted that if one of the structures of a protein in a putative protein-interaction pair is not known, it is impossible to predict whether the proteins will interact with each other. Not all proteins have known structures, limiting the set of protein-protein interactions that can be predicted. Predicting the structure of a protein given its sequence is a difficult problem. There are also almost yearly competitions to determine which method is the state of the art at predicting protein structures – see Lattman (2005) for the last set of published proceedings, and their website at `http://predictioncenter.org/`. We will not cover these methods in detail here as the focus of this thesis is on using sequence information.

2.4.2 Exploiting evolutionary information

When two proteins interact, it affects how they evolve as mutations in the proteins are likely to break the interaction. Breaking the interaction affects some aspect of the behaviour of the cell, and thus will generally be detrimental to its survival. Hence it is more likely that these proteins will undergo co-evolution, where either mutations are selected against, or there are compensating mutations on both interaction partners. Goh et al. (2000) investigated this co-

evolution effect on ligands and receptors. The authors first tracked the co-evolution of domains on a single protein. Using a large multiple-sequence alignment of the two interacting domains for many species, they built the corresponding phylogenetic trees, which were found to have a correlation of 0.79. This is an upper bound of how much information can be deduced from the co-evolution of the domains. When the analysis was repeated looking at the co-evolution of ligands and the receptors, the correlation was found to be 0.57, significantly less but still informative.

This correlated evolution between binding partners was exploited by Ramani and Marcotte (2003) to predict protein interactions. The task was to predict interactions between ligands and receptors, where a ligand binds to a receptor in order to have some biochemical effect upon it. In particular, they looked at hisitidine kinases and their corresponding sensors, a two-component sensor network. Ramani and Marcotte (2003) attempted to match up the ligands and receptors between a family of ligands and a family of receptors. This is because the more similar the evolution of a pair of ligand and receptors is, the more likely it is that they interact. First, the authors aligned a set of ligand sequences. This alignment was then used to generate a matrix of the evolutionary distance between each pair of ligands. Similarly they aligned a set of receptor sequences and built the corresponding matrix of the evolutionary distances from every receptor to every receptor. They then used simulated annealing to try and find the optimal permutation of the ligand matrix that most reduces the Root Mean Square (RMS) error between the two matrices. The proteins that headed each column in the respective distance matrices were then predicted to interact.

However, their method is highly computationally expensive due to having to find the best permutation of a large matrix – the number of permutations is $n!$, where n is the number of ligands and receptors. This search problem grows very quickly with the number of sequences, limiting their method to small numbers of sequences. For instance, they failed to the optimal permutation with 14 sequences, where the size of the search space was $14! \approx 10^{11}$. A more sophisticated optimisation strategy may let the method tackle larger families of ligands and receptors. Alternatively, a probabilistic specification of the problem could give posterior distributions over the quantities of interest, and remove the focus on finding a single optimal permutation.

Alternatively, it is possible to use prior knowledge to reduce the size of the search space. Jothi et al. (2005) significantly reduced the computational cost and improved the performance of this method by taking into account the phylogenetic tree of the ligands and the phylogenetic tree of the receptors. They started by pruning internal nodes that were poorly supported in either tree from both trees. The support for each node was calculated by bootstrapping. Given these pruned trees, they show that the reduction in the search space is $2^{I(T)}$, where $I(T) = \log(N!) - \log\tau(T)$ is the information content of T (the tree), N is the number of leaf nodes and

$\tau(T)$ is the number of automorphisms of T. An automorphism is an isomorphism of a graph to itself, and an isomorphism is a map where all edges between vertices are preserved between the graphs. If the phylogenetic trees are so poorly supported that all internal nodes are pruned, the search space becomes as large as with the matrix method of Ramani and Marcotte (2003). For real life examples, the reduction of the computational cost was found to be very substantial. This reduction in the computational cost improved the performance of the method as superior optimums could be found.

Species can both gain and lose genes, and functionally linked genes will tend to show correlated presences or absences across the species. This is because when one gene in the linkage is disabled or disappears, there is less selective pressure to keep the other gene. However, simply looking at the correlations treats all the species as equally diverged from each other, which ignores that some pairs of species will be more closely related. Barker and Pagel (2005) incorporated the common ancestry of the species by building a phylogenetic tree of representative sequences from the species, allowing inference of whether the common ancestors of those species contained those genes, and thus how many lose/gain events would be involved. These lose and gain events were found to be better predictors of functional linkage between the genes. Hence taking into account the phylogenetic tree of the species improved the performance of predicting these functional linkages between these genes. These functional linkages between genes can indicate that the corresponding proteins interact.

Other types of evolutionary information can be informative for predicting functional linkages between genes, and thus probable protein-interactions. For instance, Enright et al. (1999) predicted functional linkages between genes, and thus probable protein-protein interactions, by identifying gene-fusion events in complete genomes. Their underlying assumption was that if a composite protein is uniquely similar to two component proteins in another species, the component proteins are more likely to interact. While this is not a common event, it is an illustrative example of an *in silico* that takes into account the genomic context. See also von Mering et al. (2003), who combined a variety of such predictors that take into account the genomic context into a single predictor of functional linkages between genes.

2.5 Chapter conclusion

In this section, we have reviewed methods for discovering sequence motifs, and a variety of methods for predicting protein-protein interactions, focusing on methods that predict these interactions as the result of domain-domain interactions. However, the models for locating motifs can be too simplistic to pick up the differences between apparently similar motifs, which will tend to be merged into a single motif instead. The differences between these motifs can only be judged to be biologically significant with extra biological information. In the next

chapter, we will introduce a discriminative method to tease apart the short peptide motifs that describe the binding sites of the various SH3 domain by building a model of the resulting protein-protein interactions.

Chapter 3

A regularised discriminative model for predicting protein-peptide interactions

- Parts of this chapter have been published as Lehrach et al. (2006a), submitted in 2005.

3.1 Chapter Introduction

In the last chapter, we reviewed the current literature about protein-protein interaction prediction and motif discovery methods. In this chapter, we will propose a novel discriminative method that can distinguish between similar binding site motifs based on their interactions.

3.2 Chapter abstract

Short well-defined domains known as peptide recognition modules (PRMs) regulate many important protein-protein interactions involved in the formation of macromolecular complexes and biochemical pathways. Since high-throughput experiments like yeast two-hybrid and phage display are expensive and intrinsically noisy (see for instance Ito et al., 2001), it would be desirable to more specifically target or partially bypass them with complementary in silico approaches. In the present chapter, we present a probabilistic discriminative approach to predicting PRM-mediated protein-protein interactions from sequence data. The model is motivated by the discriminative model of Segal and Sharan (2004) as an alternative to the generative approach of Reiss and Schwikowski (2004). In our evaluation, we focus on predicting the interaction network. As proposed by Williams (1995), we overcome the problem of susceptibility to over-fitting by adopting a Bayesian Maximum A Posteriori (MAP) approach based on a Laplacian prior in parameter space.

The proposed method was tested on two datasets of protein-protein interactions involving 28 SH3 domain proteins in *Saccharmomyces cerevisiae*, where the datasets were obtained with

different experimental techniques. The predictions were evaluated with an out-of-sample receiver operator characteristic (ROC) curves. In both cases, Laplacian regularisation turned out to be crucial for achieving a reasonable generalisation performance. The Laplacian-regularised discriminative model outperformed the generative model of Reiss and Schwikowski in terms of the area under the ROC curve on both datasets. The performance was further improved with a hybrid approach, in which our model was initialised with the motifs obtained with the method of Reiss and Schwikowski.

3.3 Introduction

Peptide recognition modules (PRMs) are specialised compact protein domains that mediate many important protein-protein interactions. They are responsible for the assembly of critical macromolecular complexes and biochemical pathways (Pawson and Scott, 1997), and they have been implicated in carcinogenesis and various other human diseases (Sudol and Hunter, 2000). PRMs recognise and bind to peptide ligands that contain a certain class of proline rich motif. This motif tends to fold to a conserved structure known as the polyproline type II (PPII) helix. See Li (2005) for a review of the current literature about the PRMs. One of the most actively studied PRMs is the SH3 domain, which binds to peptide ligands that contain a particular type of proline-rich core.

Tong et al. (2002) carried out two extensive experimental studies to infer the network of SH3-mediated protein-protein interactions in *Saccharmomyces cerevisiae*. They identified 28 SH3 domain proteins in the *S. cerevisiae* proteome, which were used as baits and screened against conventional and Proline-rich libraries in a yeast two-hybrid (Y2H) experiment (Twyman, 2004). In a second independent study, they screened random peptide libraries by phage display (Twyman, 2004) to identify the consensus sequence for preferred ligands that bind to each PRM. Based on these consensus sequences, they inferred a protein-protein interaction network that links each PRM to proteins containing the preferred ligand. Since both experimental procedures are intrinsically noisy, the two independently inferred interaction networks were found to show only a modest degree of overlap.

Reiss and Schwikowski (2004) addressed the question of whether computational *in silico* approaches would allow some of the difficult and expensive experimental procedures to be more specifically targeted, or even bypassed altogether. To this end, they developed a probabilistic generative model of the SH3 ligand peptides, based on the widely used Gibbs sampling motif finding algorithm (Lawrence et al., 1993; Liu et al., 1995). Directly applying the standard Gibbs motif sampler to the *S. cerevisiae* SH3 interaction data faces the difficulty that each SH3 domain is only involved in a small number of interactions (between 1 and 20), which leads to a poor motif conservation and a high susceptibility to random artifacts due to the small sample

size. Conversely, searching for a single motif in all identified SH3 domains lacks the specificity to identify anything but a broad consensus pattern. Reiss and Schwikowski (2004) therefore devised a compromise strategy, where the network information was used as a prior on the structure of individual motifs, which were searched for with a modified version of the Gibbs motif sampler. The prior was adjusted to become discriminative, giving higher probability to those motifs that are distinct from non-binding motifs.

Reiss and Schwikowski (2004) encouragingly demonstrate that a probabilistic model trained on protein sequences and observed physical interactions can succeed in independently predicting new protein-protein interactions mediated by SH3 domains. However, a shortcoming of their model is a dependence on tuning parameters that have to be chosen in advance by the user and that are not inferred from the data. These state how similar the binding sites moifs of the different SH3 domains are to each other and how strongly to attempt to discriminate between the binding sites of different SH3 domain – see Section 3.4.1 and in particular Equation (3.3) for details. Inappropriate values reduce the performance of their algorithm to using standard motif searching algorithms, and it is unlikely that universal values applicable to different protein (super-) families exist. Also, the proposed model borrows substantial strength from its heuristic discriminative modification of the prior, which again depends on various tuning parameters.

This chapter proposes an alternative *in silico* method for the prediction of SH3-mediated protein-protein interactions, which addresses some of the shortcomings of the model introduced by Reiss and Schwikowski (2004). A key feature of our model is that it is discriminative: given a set of protein sequences, the model only attempts to find domains that distinguish between different SH3 binding domains. This is in contrast to the approach of Reiss and Schwikowski (2004), which is based on a generative model of the whole sequence. As discussed in Segal and Sharan (2004), a generative approach can be confounded by repetitive or over-represented motifs that are unrelated to PRM-peptide interactions, which our discriminative model avoids by formulating the learning problem in terms of a supervised classification problem.

The model we propose is based on a DNA-sequence model applied by Segal et al. (2002) and Segal and Sharan (2004). However, due to the larger size of the alphabet (20 amino acids instead of 4 nucleotides) and the small number of interactions per SH3 domain, their maximum likelihood approach to parameter estimation is bound to lead to serious over-fitting. An essential component of our approach, therefore, is the inclusion of a regularisation scheme, resulting in a maximum a posteriori (MAP) or penalised maximum likelihood estimate of the parameters.

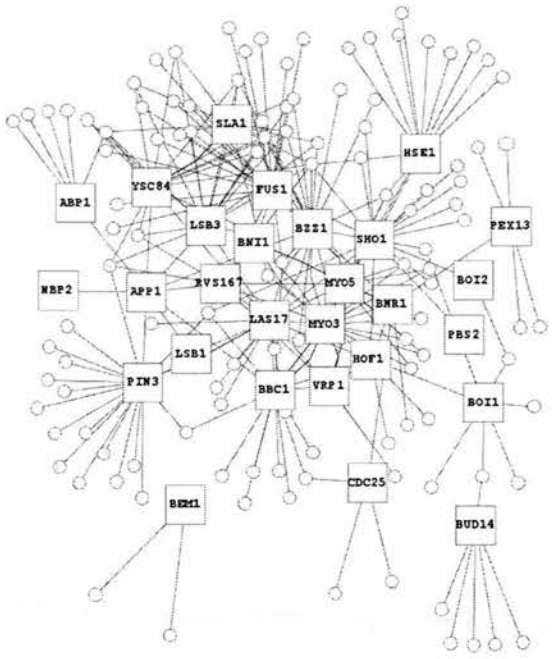


Figure 3.1: The yeast two hybrid interaction network of the SH3 domains in yeast. The labelled squares represent the central SH3 domains, while the circles represent the peripheral proteins that were found to bind to the SH3 domains.

3.4 Methods

In this section, we first define the problem, followed by an overview of the model of Reiss and Schwikowski (2004). We then derive our discriminative model and describe how to apply it. Table 3.4 summarises our notation. Let $\mathbf{D} = \{d_i\}$ denote a set of SH3 domains, and $\mathbf{S} = \{s_j\}$ a set of protein sequences. We introduce a binary variable $\varepsilon_{ij} \in \{0,1\}$, where $\varepsilon_{ij} = 1$ indicates that sequence s_j binds to SH3 domain d_i, while $\varepsilon_{ij} = 0$ indicates the absence of an interaction. We assume that we are given a protein interaction network $\mathbf{E} = \{\varepsilon_{ij}, d_i \in \mathbf{D}, s_j \in \mathbf{S}\}$ from a Y2H or phage display experiment. The objective is to derive a model that predicts this network from the sequences alone.

3.4.1 The generative model of Reiss and Schwikowski

Reiss and Schwikowski (2004) model $P(s_j|\varepsilon_{i,j} = 1)$, the probability of the sequence s_j given that it binds to the PRM d_i. The PRM for a domain d_i is modelled as a position specific scoring matrix (PSSM) $\Theta_i = \{\theta_{i,k,l}\}$, where $\theta_{i,k,l} \in [0,1]$ is the probability of observing amino acid l in the k^{th} position of the i^{th} PSSM (that is, the PSSM that indicates binding to the PRM d_i). $\mathbf{\Theta} = \{\Theta_i\}$ is the set of all PSSMs. Each position in the PRM is modelled as an independent discrete distribution - in other words, for all d_i, for all positions k, $\sum_{l=1}^{20} \theta_{i,k,l} = 1$

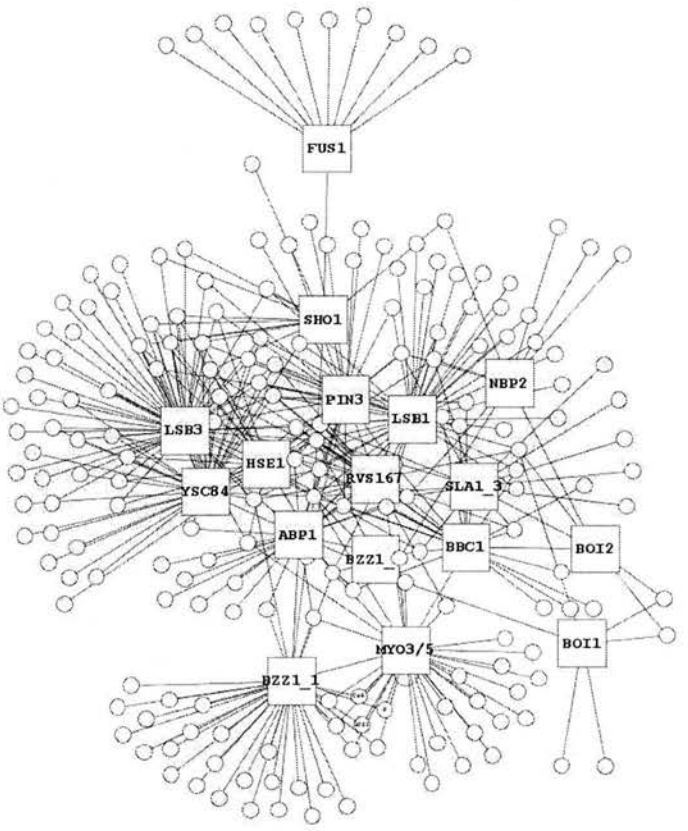


Figure 3.2: The phage display interaction network of the SH3 domain in yeast, laid out in an identical fashion to Figure 3.1.

holds. They also model the background distribution as a zeroth order Markov model $\theta_{0,l}$, where again $\sum_{l=1}^{20} \theta_{0,l} = 1$.

Given there is an interaction between domain d_i and sequence s_j, they introduce a hidden variable $a_{i,j}$, where $a_{i,j}+1$ indicates the position of the first binding site of the binding motif in sequence s_j. Note that a_{ij} ranges from 0 to $n_j - p$, where n_j is the length of the j^{th} sequence s_j and p is the length of the binding motif. $\mathbf{A} = \{a_{ij}\}$ is the set of all hidden location variables. The residues involved in the binding are then modelled as:

$$P\left(s_{j,a_{i,j}+1}, s_{j,a_{i,j}+2}, \ldots, s_{j,a_{i,j}+p} | \Theta_i, \varepsilon_{i,j} = 1\right) \quad = \quad \prod_{k=1}^{p} \theta_{i,k,s_{j,k+a_{i,j}}}. \tag{3.1}$$

The likelihood of sequence s_j with binding events $\mathbf{E}_{.,j}$ to domains $\mathbf{D} = \{d_i\}$ (with PSSMs $\mathbf{\Theta}_{.,j}$) at the corresponding binding sites $\mathbf{A}_{.,j}$ may then be written as (see Reiss and Schwikowski, 2004):

$$P(s_j, \mathbf{E}_{.,j}, \mathbf{A}_{.,j} | \mathbf{\Theta}, \theta_0) \quad \propto \quad \prod_{q=1}^{n_j} \theta_{0,s_{j,q}} \prod_{i=1}^{|\mathbf{E}_{.j}|} \left(\prod_{k=1}^{p} \frac{\theta_{i,k,s_{j,k+a_{i,j}}}}{\theta_{0,s_{j,k+a_{i,j}}}} \right)^{\varepsilon_{i,j}}. \tag{3.2}$$

The Gibbs motif sampler works by sampling the location parameters $\{a_{i,j}\}$ and the PSSM pa-

rameters $\{\Theta_i\}$ from the posterior distribution with Gibbs sampling, iterating between sampling $\{\Theta_i\}$ given $\{a_{i,j}\}$ and then $\{a_{i,j}\}$ given $\{\Theta_i\}$. The posterior distributions depend on the data via sufficient statistics that are summarised in the matrices $C_{i,j}$, whose elements are defined as $C_{i,j,k,l} = \delta(s_{j,a_{i,j}+k} = l)$ where δ is the indicator function. In words: $C_{i,j,k,l}$ is 1 if the k^{th} position of the binding motif in sequence s_j that binds to PRM domain d_i is amino acid l. Otherwise, it is zero. As opposed to the standard Gibbs sampler, Reiss and Schwikowski (2004) made use of the protein-protein interaction information $\mathbf{E} = \{\varepsilon_{ij}\}$ in computing the modified sufficient statistics $\tilde{C}_{i,j}$, which they define as follows:

$$\tilde{C}_{i,j} = \sum_a \varepsilon_{a,j} C_{a,j} + p_0 \sum_a \sum_b \varepsilon_{a,b} C_{a,b} + p_1 \sum_b \varepsilon_{i,b} C_{i,b}. \tag{3.3}$$

The third term encourages similarity of the binding motifs in sequences that bind to the same PRM domain. The second term encourages all binding motifs of all SH3 domains to be similar. The first term increases the similarity between the binding motifs for SH3 domains linked by 'promiscuous' sequences, that is, those sequences binding to more than one SH3 domain. While this approach is intuitively appealing, the scheme depends on two tunable parameters p_0 and p_1, which have to be set by the user in advance. Furthermore, the counting matrices are adjusted in a discriminative way: when two sites equally match a given PSSM Θ_i, then the one that is most dissimilar to a third, highly conserved but non-binding PSSM $\hat{\Theta}_i$, should preferentially be chosen. Define $\hat{P}(a_{i,j})$ to be proportional to a rank test between probabilities of all SH3 domains that are predicted to bind to any site on that sequence binding to that site against the probabilities of all SH3 domains which are not predicted to bind to any site on that sequence of binding to that site. This then favours sites which are different from the binding site motifs of the non-binding SH3 domains. These binding site probabilities are then adjusted with user-tunable parameter q_d, and ultimately defined to be $P(a_{i,j}) \propto \hat{P}(a_{i,j}) + q_d$. Hence, this discriminative prior depends on the choice of the rank test and the user-tunable parameter q_d, which has to be set in advanced by the user in the same manner as p_0 and p_1.

3.4.2 A discriminative model

In our discriminative approach, we do not directly model θ_0, the background distribution and Θ_i, the motif distributions. Instead, we directly model the probability of the occurrence of a binding motif for the i^{th} PRM in sequence s_j: $P(\varepsilon_{i,j}|s_j)$. We start with a very similar model to Reiss and Schwikowski (2004). The probability of a sequence given a binding site motif in position m is shown in Equation (3.4). Equation (3.5) shows the probability of a sequence without a motif. In order to keep the notation concise, we drop the conditioning on $\mathbf{\Theta}_i$ and θ_0

from this derivation.

$$P(s_j|\varepsilon_{i,j}=1,a_{i,j}=m) = \prod_{q=1}^{n_j}\theta_{0,s_{j,q}}\prod_{k=1}^{p}\frac{\theta_{i,k,s_{j,m+k}}}{\theta_{0,s_{j,m+k}}}. \quad (3.4)$$

$$P(s_j|\varepsilon_{i,j}=0) = \prod_{q=1}^{n_j}\theta_{0,s_{j,q}}. \quad (3.5)$$

We assume a uniform prior over $a_{i,j}$, the possible binding positions for the binding site motif of the SH3 domain:

$$P(a_{ij}=m) = \frac{1}{n_j-p+1}. \quad (3.6)$$

Marginalising out the unknown binding position $a_{i,j}$ gives:

$$P(s_j|\varepsilon_{i,j}=1) = \prod_{q=1}^{n_j}\theta_{0,s_{j,q}}\frac{1}{n_j-p+1}\sum_{m=0}^{n_j-p}\prod_{k=1}^{p}\frac{\theta_{i,k,s_{j,m+k}}}{\theta_{0,s_{j,m+k}}}. \quad (3.7)$$

Applying Bayes' rule:

$$P(\varepsilon_{i,j}=1|s_j) = \frac{P(s_j|\varepsilon_{i,j}=1)P(\varepsilon_{i,j}=1)}{P(s_j)}, \quad (3.8)$$

where $P(s_j)=\sum_{\varepsilon_{i,j}=0}^{1}P(s_j|\varepsilon_{i,j})P(\varepsilon_{i,j})$, and combining Equations (3.5) and (3.7), we get:

$$P(\varepsilon_{i,j}=1|s_j,\mathbf{W},\mathbf{T}) = \text{logit}\left(\log\left(\frac{1}{n_j-p+1}\sum_{m=0}^{n_j-p}\exp\left\{\sum_{k=1}^{p}W_{i,k,s_{j,m+k}}\right\}\right)+T_i\right). \quad (3.9)$$

where we have defined $W_{i,k,l}=\log\frac{\theta_{i,k,l}}{\theta_{0,l}}$, $T_i=\log\frac{P(\varepsilon_{i,j}=1)}{P(\varepsilon_{i,j}=0)}$ and $\text{logit}(x)=\left(1+e^{-X}\right)^{-1}$. The dependence on $\boldsymbol{\Theta}_i$ and θ_0 has been replaced by a dependence on the weights $\mathbf{W}=\{W_{i,k,l}\}$. For convenience, we also introduce a dependence on the set of thresholds $\mathbf{T}=\{T_i\}$, replacing our earlier implicit dependence on our choice of $P(\varepsilon_{i,j}=1)$. We can now apply this discriminative model, which corresponds to Equation (2) in Segal et al. (2002), and the Equation in Section 2.1 of Segal and Sharan (2004), to infer the presence of an interaction between a peptide sequence and an SH3 domain.

3.4.3 Parameter estimation

Having specified the model, we next need to estimate the parameters, which are the set of weights $\mathbf{W}=\{W_{i,k,l}\}$ and thresholds $\mathbf{T}=\{T_i\}$. A standard way to optimise these parameters, adopted for instance in Segal et al. (2002) and Segal and Sharan (2004), is to follow a maximum likelihood approach. Given the training data D, which is the set of all training sequences s_j and binding interaction indicator variables $\varepsilon_{i,j}$, we want to maximise the log likelihood $-E_D$:

$$\begin{aligned}-E_D &= \log P(D|\mathbf{W},\mathbf{T})\\ &= \sum_{i,j}\ \varepsilon_{i,j}\ \log P(\varepsilon_{i,j}=1|s_j,\mathbf{W},\mathbf{T})+\\ &\qquad (1-\varepsilon_{i,j})\ \log\left(1-P(\varepsilon_{i,j}=1|s_j,\mathbf{W},\mathbf{T})\right).\end{aligned} \quad (3.10)$$

Note that $P(\varepsilon_{i,j} = 1|s_j, \mathbf{W}, \mathbf{T})$ has been defined in Equation (3.9). It is straightforward to derive the partial derivatives $\frac{\partial E_D}{\partial W_{i,k,l}}$ and $\frac{\partial E_D}{\partial T_i}$, which allows us to apply an iterative gradient-based optimisation scheme.

3.4.4 Regularisation

A shortcoming of the maximum likelihood approach discussed in the previous section is its susceptibility to over-fitting as for each SH3 domain, we have $20p+1$ parameters to estimate. This exceeds the number of peptide sequences an SH3 domain binds to and hence calls for the implementation of an effective regularisation scheme. To be more specific, there are 181 parameters per SH3 domain, and the mean number of peptide sequences each SH3 domain binds to in the yeast two-hybrid dataset is 10. A standard approach widely applied in machine learning is to impose a prior probability on the weights $\mathbf{W}$ such that large weight values are discouraged and *a priori* a value of zero is assumed. Define E_W to denote a function of $\mathbf{W}$ that is monotonically increasing with the magnitude of the weights $\mathbf{W}$. We define the prior:

$$P(\mathbf{W}|\alpha) = \frac{1}{Z}\prod \exp(-\alpha E_W(\mathbf{W})), \tag{3.11}$$

where Z is a normalisation constant, and α represents a scale factor. This choice of prior is particularly meaningful in our application. From the definition of the weights in the text below Equation (3.9) it is seen that $W_{i,k,\cdot} = 0$ corresponds to the assumption that the amino acid distribution at the k^{th} position of the i^{th} motif, $\theta_{i,k,\cdot}$, is equal to the background distribution $\theta_{0,\cdot}$. Consequently, the l^{th} amino acid occurring in the k^{th} motif position provides no information about whether the amino acid is part of the background or part of the motif, which considering the larger number of parameters compared to sequences will be a common occurrence.

A prior commonly used in machine learning is the Gaussian distribution for $P(\mathbf{W}|\alpha)$ (see, for instance, MacKay (1992)), where:

$$E_W(\mathbf{W}) = \frac{1}{2}\sum_{i,k,l} W_{i,k,l}^2. \tag{3.12}$$

Less widely applied, but for our application particularly appropriate, is the Laplacian prior (Williams, 1995):

$$E_W(\mathbf{W}) = \sum_{i,k,l} |W_{i,k,l}|. \tag{3.13}$$

The difference between these priors will be explained shortly. Note that we have left the thresholds $\mathbf{T}$ unregularised, as suggested in Williams, 1995. This corresponds to a uniform prior

$$P(\mathbf{T}) = \text{Constant}. \tag{3.14}$$

From our definition of T_i from below Equation (3.9), this is seen to correspond to a lack of prior knowledge about the global connectivities of the different SH3 domains. If this knowledge is

available, it is straightforward to replace $P(\mathbf{T})$ by a more informative prior. Now, the ideal approach would be to follow a fully Bayesian approach and sample the parameters $[\mathbf{W},\mathbf{T}]$ and the so-called hyperparameter α from the posterior distribution $P(\mathbf{W},\mathbf{T},\alpha|D)$. Since this distribution is not available in closed form and cannot directly be sampled from, we have to resort to a numerical approximation with Markov chain Monte Carlo. Neal (1996) has applied this scheme in the context of neural networks, where he observed a significant improvement in the generalisation performance over maximum likelihood. Originally, we thought that the computational costs of inferring so many parameters would be excessive and hence we investigated the approach outlined in this chapter. Note that in Section 3.9.1 we describe an alternative sampling scheme which could improve the performance of the model described in this chapter, and we, later in this section, reduce this coupling between the parameters making a sampling approach easier. In our alternative approach, rather than sample from the posterior distribution, we find the parameters that maximise this distribution, the so-called *maximum a posteriori* (MAP) estimate:

$$[\mathbf{W},\mathbf{T}]_{opt} = \arg\max_{\mathbf{W},\mathbf{T}} P(\mathbf{W},\mathbf{T}|D,\alpha). \tag{3.15}$$

Now:

$$P(\mathbf{W},\mathbf{T}|D,\alpha) \propto P(D|\mathbf{W},\mathbf{T})P(\mathbf{W}|\alpha)P(\mathbf{T}). \tag{3.16}$$

From Equations (3.10), (3.11), and (3.14) we see that the optimisation of the thresholds $\mathbf{T}$ remains unaffected by the proposed regularisation scheme. For the weights $\mathbf{W}$, however, we now have to minimise the modified cost function

$$E^*(\mathbf{W}) = E_D(\mathbf{W}) + \alpha E_W(\mathbf{W}). \tag{3.17}$$

where we have two alternative functions for E_W; see Equations (3.12) and (3.13). These two priors can be justified in a maximum entropy sense under different invariance assumptions, as discussed in Williams (1995). The practical difference between the two priors can be understood from the derivatives of the regularisation term E_W. In the Gaussian case, this derivative is proportional to the size of the weights: $\left|\frac{\partial E_W}{\partial W_{i,k,l}}\right| = |W_{i,k,l}|$. This implies that large weights are more heavily penalised than small weights, and the model tends to end up with a large number of small weights. For the Laplacian prior, the derivative is constant: $\frac{\partial E_W}{\partial W_{i,k,l}} \propto 1$. This imposes less of a penalty on large weights, while driving small weights more strongly down to zero. In fact, the discontinuity of the derivative at the origin $W_{i,k,l} = 0$ can be used for a pruning scheme, as discussed in Williams (1995).

The proposed regularisation scheme seems to depend on the hyperparameter α. In fact, this hyperparameter can be integrated out:

$$P(\mathbf{W}) = \int P(\mathbf{W}|\alpha)P(\alpha)d\alpha. \tag{3.18}$$

GEN	Generative model of Reiss and Schwikowski (2004)
DIS-I	Discriminative model, informative initialisation
DIS-E	Ensemble of discriminative models, random initialisations.

Table 3.1: An overview of the models compared in our study.

Since α is a scale parameter, it is reasonable to use the improper $1/\alpha$ ignorance prior (Williams, 1995). It is then straightforward to show that for $E_W(\mathbf{W})$, as in Equations (3.11) and (3.13),

$$-\log P(\mathbf{W}) = |\mathbf{W}| \log E_W, \tag{3.19}$$

where $|\mathbf{W}|$ is the number of weight parameters. Replacing $P(\mathbf{W}|\alpha)$ by $P(\mathbf{W})$ in Equation (3.16):

$$P(\mathbf{W},\mathbf{T}|D) \propto P(D|\mathbf{W},\mathbf{T})P(\mathbf{W})P(\mathbf{T}), \tag{3.20}$$

and noting that the threshold parameters T_i are unregularised (corresponding to a uniform prior), this leads to the following modification of the objective function E^* (compare with Equation (3.17)):

$$E^*(\mathbf{W}) = E_D(\mathbf{W}) + |\mathbf{W}| \log E_W. \tag{3.21}$$

Now, taking derivatives we get:

$$\frac{\partial E^*}{\partial W_{i,k,l}} = \frac{\partial E_D}{\partial W_{i,k,l}} + \tilde{\alpha}\frac{\partial E_W}{\partial W_{i,k,l}}, \tag{3.22}$$

where the effective hyperparameter $\tilde{\alpha} = \frac{|\mathbf{W}|}{E_W}$ is determined adaptively during training. Hence, as opposed to Reiss and Schwikowski (2004), we have no arbitrary parameters that would need to be hand-tuned by the user. As an aside, we notice that the integration over the hyperparameters has been criticised by MacKay (1999) on the grounds that in conjunction with the MAP approximation it may lead to over-regularisation. An alternative, proposed by MacKay (1992), is a computationally more expensive maximum likelihood type II optimisation of α. Interestingly, these approaches lead to identical results when using the Laplacian prior (Williams, 1995), hence rendering the MAP approach more valid than in the Gaussian case.

The regularisation method proposed in this section can easily be generalised to allow for more than one hyperparameter α. In fact, we divided the weights $W_{i,k,l}$ into separate weight groups, one for each SH3 domain protein, where each weight group was associated with a separate hyperparameter. Such weight groups have been found in previous studies to improve the generalisation performance of neural networks (MacKay, 1992). To reduce the opacity of the notation, we have not made this modification explicit in the text. This has the additional advantage of making sampling schemes easier to implement, as the parameters for each SH3 domain decouple, meaning that only 181 parameters need to be simultaneously inferred.

3.4.5 The algorithm

We adapted the parameters with conjugate gradients, using the MATLAB implementation in the NETLAB library (Nabney, 2002). We rescaled the objective function of Equation (3.10) by assuming that there is a small $\zeta = 10^{-8}$ chance of measurements being incorrect. The effect of this is to constrain the objective function to remain finite within floating point accuracy, leading to a significantly faster rate of convergence. This corresponds to the model of uncertainty in measurements discussed in Deng et al. (2002). The weights $W_{i,k,l}$ were regularised according to Equation (3.22) and we updated the effective hyperparameter $\tilde{\alpha}$ every 10 iterations, as described below Equation (3.22). After each update, the search direction of the conjugate gradients method was reset. For parameter pruning in conjunction with the Laplacian prior, we followed the procedure described in Williams (1995).

3.5 Simulations

We removed SH3 domain proteins that only bind to a single peptide, as there would be no way to validate these interactions on an independent test set. With this modification, the Y2H dataset (displayed in Figure 3.1) has 28 SH3 domains, 143 binding partners, and 285 interactions, while the phage display dataset (displayed in Figure 3.2) contains 17 SH3 domains, 207 binding partners, and 381 interactions. Further details are in the supplementary material of Tong et al. (2002).

We evaluated the generalisation performance with a 10-fold cross-validation scheme where the data was randomly partitioned into 10 folds. The generalisation performance was then evaluated on the current fold, and the other 9 folds were used for training. We obtained an average out-of-sample performance by repeating this for all 10 folds. Only the ensemble method described in Section 3.7 used random restarts.

The performance was measured in terms of ROC curves, which are obtained by subjecting the predicted posterior probabilities $P(\varepsilon_{ij}|s_j)$ to various threshold parameters $\theta \in [0,1]$. By numerically integrating over the whole parameter range $\theta \in [0,1]$ we obtain the area under the ROC curve. This so-called AUROC score ranges from 0.5 for a random predictor to 1.0 for a perfect predictor, with larger values generally indicating a better performance. Since the left part of the ROC curves, where the number of false positives is low, is often of particular interest, we also restrict the integral to false positive values of less than 0.1. We refer to the resulting score as AUROC01.

To evaluate the performance of the generative model of Reiss and Schwikowski (2004), we used the software provided by the authors, which is available from `http://sf.net/projects/netmotsa`. Recall that the generative model depends on various tuning parameters, which are not inferred from the data but rather have to be set by the user in advance.

For our comparative study, we used the default values defined in the software of Reiss and Schwikowski (2004). These parameters had been optimised by the authors on the same data set as used in our study; hence they should reflect a quasi-ideal performance.

3.6 Regularisation

3.6.1 The effect of regularisation

The effect of regularisation is clearly demonstrated in Figure 3.3. Figure 3.3a shows various graphs obtained from unregularised training simulations. The bottom left sub-figure shows the evolution of the AUROC scores during the training process when no regularisation is applied. The AUROC score on the training set increases monotonically and, in fact, converges to its maximum possible value of AUROC=1. However, on the test set, the AUROC score increases only during the first few iterations, after which it steadily deteriorates. This behaviour is also reflected in the cross-entropy (E_D in Equation (3.10)), shown in the bottom right sub-figure – note that a high AUROC score corresponds to a low cross-entropy while a low AUROC score corresponds to a large cross-entropy. The final training ROC, shown in the top right sub-figure, is that of an ideal classifier where all interactions are detected at a zero false positive rate (AUROC=1.0). However, the test ROC is very poor and not that much better than a random predictor (AUROC=0.59). This effect is called overfitting. Compare this with the results obtained from the regularised training simulations, which are shown in Figure 3.3b. The training and test ROCs are very similar, with AUROC scores of 0.66 and 0.71, respectively. These scores are markedly worse than the ideal AUROC score of 1.0. However, they show a considerable improvement compared to the final test AUROC of 0.59 that was obtained without regularisation. The effect of regularisation is also clearly seen in the evolution of the AUROC and cross-entropy scores during the training cycles (bottom graphs). The monotonic trends on the training set only prevail for the first 10-20 iterations, after which the trend is reversed until it eventually reaches a stationary plateau. The final training and test scores are similar, clearly indicating that overfitting is avoided.

The improvement of the regularisation scheme also becomes clear from the results of the ten-fold cross-validation analysis shown in Figure 3.4. Note that the AUROC scores of the Laplacian regularised model substantially exceed those of the unregularised model (Y2H: 0.66 vs 0.60, PD: 0.71 vs 0.57).

3.6.2 The relative performance of Gaussian and Laplacian regularisation

A comparison of the performance of the different regularisation schemes is shown in Figure 3.4. For both the phage display and Y2H networks, the ROC curves obtained with the Gaussian-

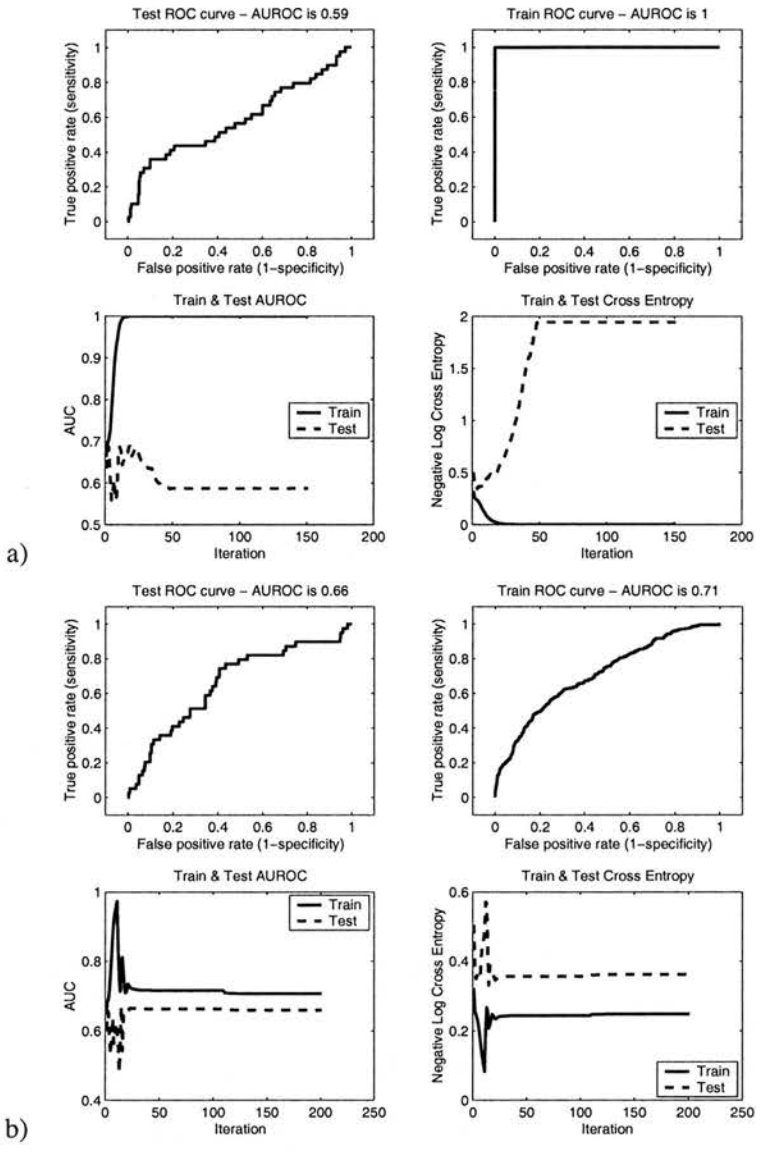


Figure 3.3: An illustration of the importance of regularisation, demonstrated on the phage display network. Sub-figure a) shows the performance of the predictor in an unregularised framework, while Sub-figure b) shows the equivalent simulation performed with a Laplacian regulariser. In both Sub-figures a) and b), the top left graph represents the final ROC curve obtained on the test data, while the upper right graph shows the corresponding ROC curve for the training data. The bottom left graph shows how the AUROC score changes as the training algorithm is run. The bottom right graph shows how the negative log cross entropy of Equation (3.10) changes during the run, where lower values indicate better classifiers. Comparing the test and training ROC curves and AUROC scores in Sub-figure a) shows that the unregularised model is clearly overfitting as the training performance is perfect while the test performance is poor. In Sub-figure b), however, there is a very good correspondence between the test and the training scores, indicating that the regulariser is working as intended and giving us an improved final test performance.

		GEN	DIS-I	DIS-E
Yeast two-hybrid	AUROC	0.61	0.67	0.67
	AUROC01	0.17	0.17	0.16
Phage display	AUROC	0.69	0.83	0.71
	AUROC01	0.17	0.44	0.19

Table 3.2: AUROC and AUROC01 scores obtained with ten-fold cross-validation for different models on the yeast two-hybrid and phage display data. The corresponding ROC curves are shown in Figure 3.5.

regularised discriminative approach are very similar to those of the naïve method, where:

$$P(\varepsilon_{i,j} = 1) = \frac{1}{|\mathbf{S}|} \sum_{j=1}^{|\mathbf{S}|} \varepsilon_{i,j}. \tag{3.23}$$

The naïve method predicts the probability of an interaction as the frequency of the SH3 domain interacting with the peptide sequences. The performance of the Gaussian-regularised discriminative approach is close to that of the naïve predictor, which suggests that the quadratic regulariser is too restrictive and pulls all weights $W_{i,k,l}$ to values close to zero. The prediction of an interaction according to Equation (3.9) is therefore dominated by the threshold parameters T_i, which are unregularised. These thresholds represent the prior probability of an interaction, as obtained from the overall connectivity of a node; see the definition of T_i below Equation (3.9). Hence, the Gaussian-regularised discriminative model effectively reduces to the naïve predictor of Equation (3.23), which explains the similarity between their ROC curves.

The Laplacian-regularised discriminative model, on the other hand, leads to ROC curves that are significantly better than those obtained with both the naïve and the unregularised discriminative model. This finding suggests that the Laplacian regulariser prevents the overfitting of the unregularised approach while avoiding the over-regularisation of the Gaussian regulariser. In particular, the improvement of the ROC curves over those of the naïve method indicates that the prediction does not only depend on the global connectivities captured by the threshold parameters T_i, as for the Gaussian regularised model, but that the weights $W_{i,k,l}$ have encoded useful additional information about the protein interaction sites.

3.7 Results

As we found that the Laplacian regularised model outperformed both the unregularised and the Gaussian regularised models, we now focus our investigations on using Laplacian regularisation. In the simulations reported in the following sections, we have compared three approaches: (1) the generative model of Reiss and Schwikowski (2004); (2) the proposed discriminative

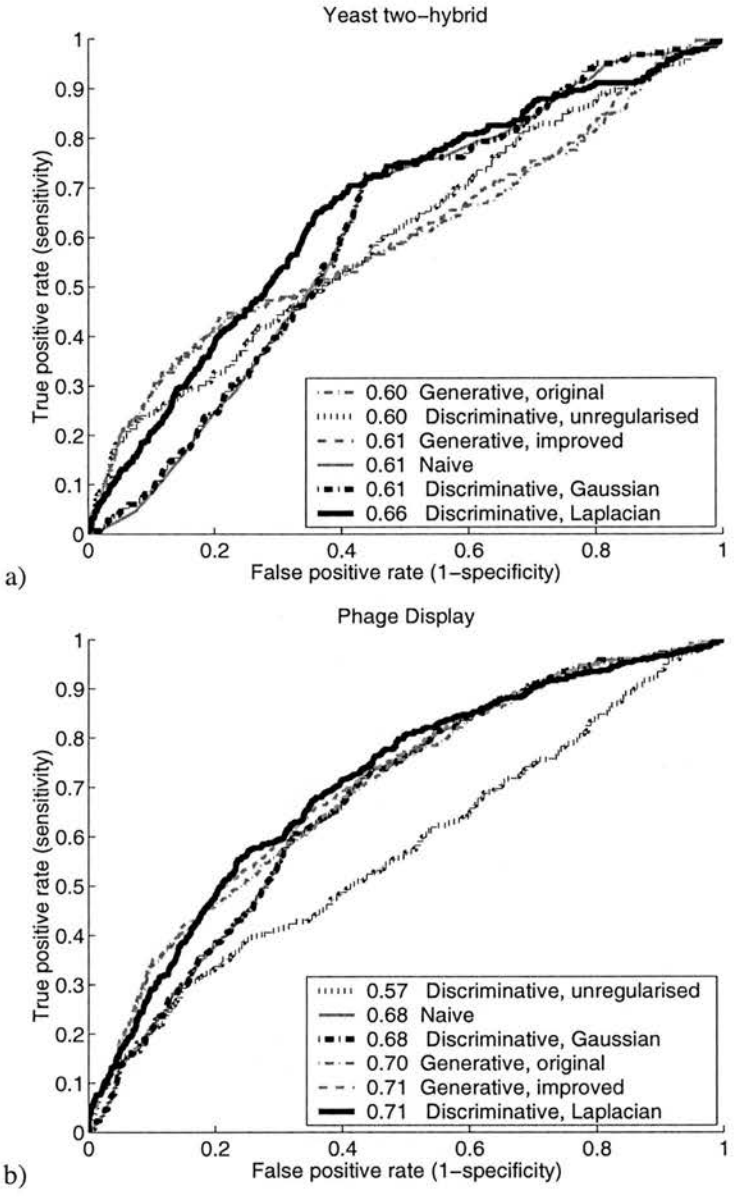


Figure 3.4: The relative performance of Laplacian and Gaussian regularisation, performed without using an ensemble method or informative initialisation. We compare the original generative model of Reiss and Schwikowski (2004), the improved generative model where the generative model instead uses the prior probabilities from Equation (3.23) and the naïve classifier from Equation (3.23) with our discriminative method. We apply three types of regularisation (none, Gaussian and Laplacian) to our discriminative method. Both Sub-figures a) and b) represent ROC curves, where the area under the ROC curve (AUROC score) is shown in the legend. Sub-figure a) shows the performance on the yeast two-hybrid (Y2H) data, while Sub-figure b) shows the performance on the phage display data.

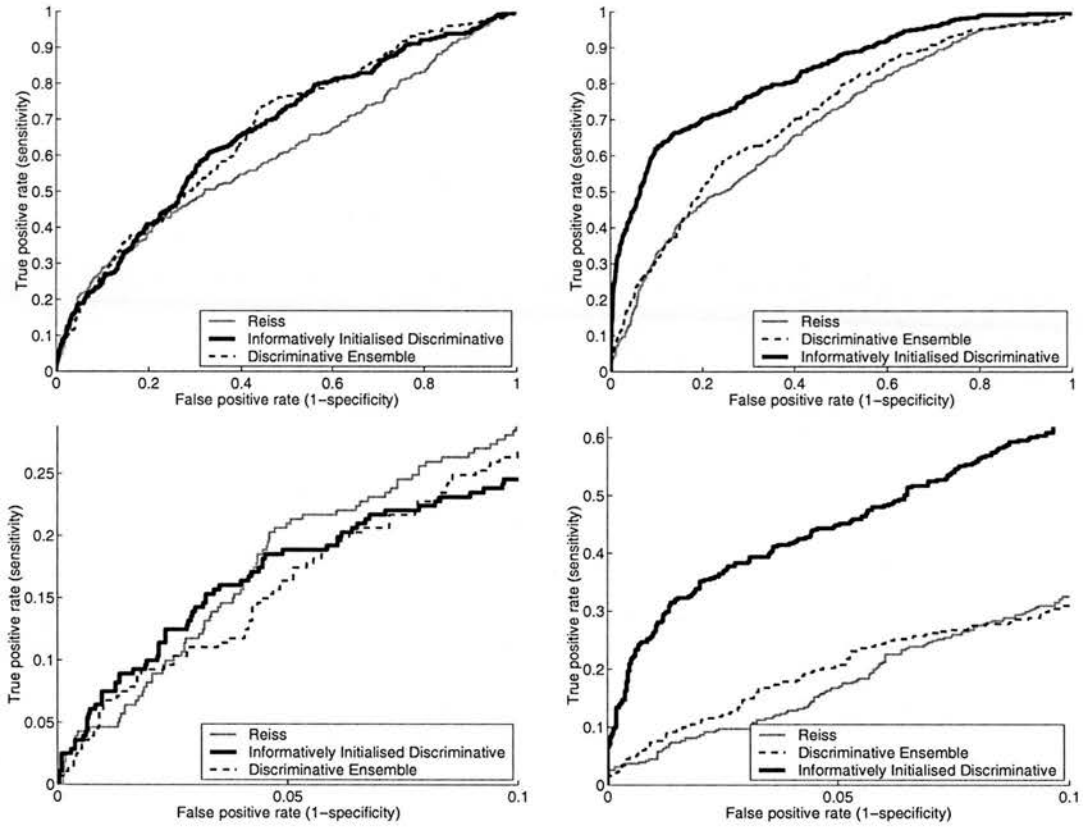


Figure 3.5: ROC curves obtained for the three methods compared in our study: GEN (narrow solid lines), DIS-I (thick solid line) and DIS-E (dashed line). A ten-fold cross-validation scheme was applied, as described in Section 3.7. *Left panel:* yeast two-hybrid network; *right panel:* phage display network. The bottom panel shows the left part of the ROC curves in a higher resolution. The corresponding AUROC scores are shown in Table 3.2.

model, where the weights were initialised from the PSSMs learnt with the generative model; and (3) an ensemble of discriminative models; this ensemble was created by training ten models from different initialisations[1], and keeping the five models with the highest training set scores. In what follows, we will refer to these methods as (1) GEN, (2) DIS-I, and (3) DIS-E, respectively; see Table 3.1 for a summary. For training the discriminative models, the Laplacian regularisation scheme was applied throughout. Note that when adapting the weights which were initialised with the PSSMs learnt with the generative model (DIS-I), an ensemble of the resulting weights was not created.

3.7.1 Assessing the prediction performance

The top left panel of Figure 3.5 shows the ROC curves obtained for the yeast two-hybrid network. Both discriminative methods, DIS-I and DIS-E, clearly outperform GEN in the right part of the graph, for false positive rates (FPR) greater than 0.3. This is reflected in higher overall AUROC scores, as seen from Table 3.2. In the left part, for $FPR < 0.3$, the three methods perform more or less equally. Plotting the ROC curves for values of $FPR < 0.1$ at a higher resolution, as done in the bottom left panel of Figure 3.5, reveals that DIS-I and GEN perform equally (AUROC01=0.17), and slightly better than DIS-E.

The right panel of Figure 3.5 shows the ROC curves obtained for the phage display network. The discriminative methods outperform the generative model, both in terms of overall (top panel) and left-side (bottom panel) performance. This improvement is considerably improved when starting the training simulations from an informative initialisation (DIS-I). Also, we found that the performance of all methods is consistently better for the phage display network than for the yeast two-hybrid network.

Note that we do not get confidence intervals from using cross-validation, as we combine all these predictions into a single set of predictions. This is then used to calculate the ROC curve, and hence the AUROC score. It does not hold that the mean of the AUROC scores is the same as the AUROC of the combined results, so displaying the standard deviation of the AUROC score between cross folds with the overall AUROC score would be meaningless

3.7.2 Locating binding regions

To test whether the proposed model is actually able to locate the binding sites, we focused on Las17, which can form protein complexes containing multiple SH3 domains. Tong et al.

[1] This is an approximation to a sampling scheme. Probably, integrating over all weights using a sampling scheme would give better performance – see Section 3.9.1.

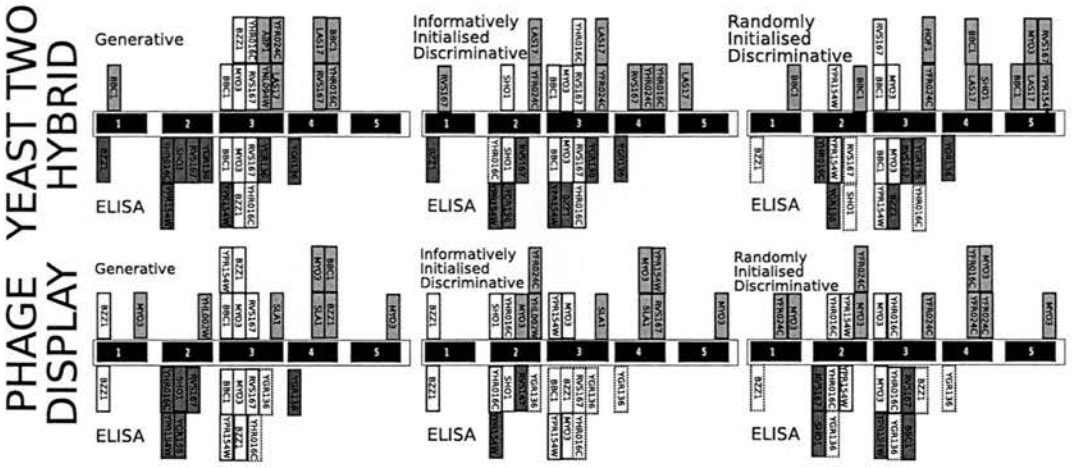


Figure 3.6: Predicting the binding locations of SH3 domain proteins interacting with Las17. The top row shows the results obtained for yeast two-hybrid; the bottom row refers to phage display. The columns correspond to three different in silico methods. Left column: the generative model of Reiss and Schwikowski (2004). Centre column: the proposed discriminative model trained from an informative initialisation. Right column: an ensemble of ten randomly initialised discriminative models. Each of the six panels represents the five proline-rich peptide fragments of Las17 studied in Tong et al. (2002). These fragments are located in the following regions: 1) 153-190, 2) 306-336, 3) 339-366, 4) 374-403, 5) 423-476. The positive interactions observed in the ELISA experiments of Tong et al. (2002) are shown in the bottom of the panel. Empty boxes indicate interactions that are predicted in silico among the 14 highest-scoring interactions. Empty boxes with a dashed border show interactions that are predicted in silico among the 28 highest-scoring interactions. Shaded boxes represent interactions that have not been predicted (false negatives). The boxes in the top of the panel refer to in silico predictions, where empty boxes indicate interactions confirmed in the ELISA experiment (true positives), and shaded boxes show predicted interactions that have not been found in the ELISA experiment (false positives).

		GEN	DIS-I	DIS-E
Yeast two-hybrid	AUROC	0.60	0.65	0.68
	p-value	0.24	0.07	0.03
Phage display	AUROC	0.77	0.84	0.71
	p-value	0.002	8×10^{-5}	0.014

Table 3.3: AUROC scores and p-values for locating binding regions in Las17. The corresponding ROC curves are shown in Figure 3.7.

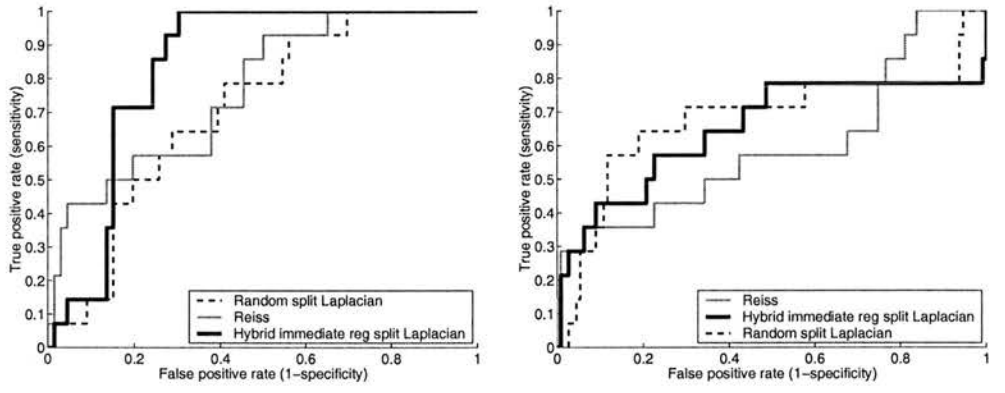


Figure 3.7: Predicting binding locations in Las17. The ROC curves capture the prediction of binding locations of SH3 domain proteins interacting with Las17. *Left panel:* Yeast two-hybrid. *Right panel:* Phage display. Each panel consists of three graphs, which refer to the three models compared in our study. *Thick solid line:* DIS-I. *Dashed line:* DIS-E. *Narrow solid line:* GEN. Further details are provided in the text.

(2002) have applied an enzyme-linked immunosorbent assay (ELISA) to identify the region of the target protein that binds the SH3 domain. They focused on five proline-rich peptide fragments of Las17, whose locations are indicated in the caption of Figure 3.6. In our study, we removed Las17 from the training set, and repeated the training simulations for both the yeast two-hybrid and the phage display data. We then tested whether the binding locations of SH3 domain proteins interacting with Las17 could be correctly predicted.

We evaluated the models in three different ways. In the first evaluation, we took the interactions of the ELISA experiment, reported in Tong et al. (2002), as true interactions. We applied the models to these segments separately, ranked the SH3 domains for each segment according to their segment-specific binding scores, and obtained the ROC curves from these rankings. The results are shown in Figure 3.7.

In the second evaluation, we selected the threshold that resulted in a total of 14 predicted SH3 domains. This number is equal to the total number of interactions detected with the ELISA experiments when omitting Yfr024 (Yfr024 binds only to a single sequence and was therefore omitted from our study for the reason discussed above). We then compared the predictions between the different models and with the ELISA experiment. The results are shown in Figure 3.6.

In the third evaluation, we tested whether the predictions obtained with the different models were significantly better than would be obtained by chance. This is similar to the procedure reported in Reiss and Schwikowski (2004), except that the authors do not provide details of how they obtained their p-values. We proceeded as follows. For each proline-rich segment, we ranked the SH3 domains according to the binding score predicted by the model. We divided

the SH3 domains into two classes: those detected as binding with the ELISA experiment, and those not detected as binding. We then applied the Wilcoxon rank sum (or Mann-Whitney) test to obtain the p-values.

The top panel of Figure 3.6 shows the thresholded predictions obtained for the yeast two-hybrid data. Both GEN and DIS-I predict 5, while DIS-E predicts 4 true positive interactions. The slightly worse performance of DIS-E can be understood from the ROC curves in the left panel of Figure 3.7, where DIS-E shows a poorer performance in the very left part of the graphs. Lowering the threshold turns out to be beneficial only for the discriminative models: DIS-I predicts 1, and DIS-E predicts 4 additional true interactions. Again, this finding is consistent with the ROC curves in Figure 3.7, where both discriminative models outperform GEN. The three models show a modest degree of complementarity. GEN fails to predict any SH3 domain protein binding to the second segment of Las17, and this failure persists even as the threshold is lowered. To the contrary, both discriminative models predict at least one SH3 protein binding to this segment, and this number increases as the threshold is lowered.

The bottom panel of Figure 3.6 shows the predictions obtained for the phage display data. Among the 14 highest-scoring interactions, GEN predicts 6 true positives, while DIS-I and DIS-E predict only 5 and 4, respectively. However, among the next 14 highest-scoring interactions, GEN only gains 2 extra true positives. DIS-E gains 5 extra true positives, and thus performs slightly better than GEN. Both methods are noticeably outperformed by DIS-I, which predicts all but two interactions. This improved performance is consistent with the ROC curves, shown in the right panel of Figure 3.7. While GEN obtains higher true positive rates in the left-most region of the graph, both discriminative models experience a considerable performance boost at a false positive rate of 0.18, and DIS-E shows the best performance overall.

The results discussed in this section are concisely summarised in Table 3.3. For the yeast two-hybrid data, both discriminative models slightly outperform the generative model: DIS-E (AUROC=0.68) > DIS-I (AUROC=0.65) > GEN (AUROC=0.60). For the phage display data, the discriminative model with the informative initialisation outperforms the generative model: DIS-I (AUROC=0.84) > GEN (AUROC=0.77) > DIS-E (AUROC=0.71). The performance on the phage display data is, overall, better than that on the yeast two-hybrid data, with all p-values significant. For the yeast two-hybrid data, only the p-values obtained for the discriminative methods would be regarded as significant.

3.7.3 Biological validation and application

An important practical application of the proposed method would be the cleaning and filtering of high-throughput interaction data. Our conjecture is that protein interactions that are assigned a higher posterior probability score in silico are more reliable than those with a lower score. We would therefore assume that interactions found with both the yeast two-hybrid and

the phage display experiment have higher posterior probability scores than those found with only one experiment. Phrased differently, we would assume that the intersection of the sets of interactions found with yeast two-hybrid and phage display shows an enrichment for higher-scoring in silico interactions. To test this conjecture, we extracted for both experiments the 400 highest-scoring interactions; this is the number of interactions detected experimentally with phage display. When training our model on the yeast two-hybrid data, we recovered 25 percent of the interactions in the intersection set, but only 8 percent of the interactions in the complementary non-intersection set. When training our model on the phage display data, we again recovered 25 percent of the interactions in the intersection set, but only 9 percent of the interactions in the non-intersection set. Hence, in both training simulations, we found that the subset of more reliable interactions (that is, those interactions found with both experimental methods), was noticeably enriched for high-scoring in silico predictions. This finding corroborates our hypothesis that the proposed in silico method could, in fact, offer a useful tool for filtering noisy high-throughput interactomic data.

3.8 Discussion

The model we propose is based on the assumption that protein interactions are mediated by short peptide segments that bind to PRM domains. This assumption is valid for the phage display data, which explains why all the models achieve a better performance here than on the yeast two-hybrid interaction network. Yeast two hybrid finds all peptide sequences present in the organism that bind to the SH3 domain. These interactions do not have to be mediated by a motif. In contrast, in the phage display dataset, short peptide fragments were tested against an SH3 domain. The peptide sequences that were found to bind were then used to create a model of the motif that binds to that domain. This motif is then searched for in all the peptide sequences in yeast. Peptide sequence containing this motif were then predicted to interact with that SH3 domain. This closely matches what the generative and discriminative motif finding models search for, hence the model mismatch is smaller. This leads to the significant increase in performance exhibited by all models on the phage display dataset.

Our simulations suggest that the randomly initialised discriminative model achieves a performance at least as good as the generative model of Reiss and Schwikowski (2004). While the AUROC01 score for the yeast two-hybrid network is slightly worse, the overall AUROC scores have noticeably improved. The discriminative model also shows some complementarity to the generative model with respect to locating the binding regions.

When initialising the discriminative model with the PSSMs predicted by the generative model, its performance further improves. With this informative initialisation, the discriminative model outperforms the generative model of Reiss and Schwikowski (2004) both in terms of

predicting protein interactions and locating binding regions. The improvement is particularly noticeable for the phage display network, where the data are more in line with the model assumptions.

Note that the model we have proposed has only been trained on proteins in the SH3 interaction networks. Hence, the objective of our approach is to predict the probability of a particular protein interaction, given that the protein is in the interaction network. In principle, it is straightforward to generalise this approach to not only distinguish between the different protein interactions, but also to predict whether the protein is in the interaction network. All that is required is to include an extra output node representing non-binding background sequences. However, the inclusion of background sequences, which substantially outnumber the binding sequences, would substantially increase the computational costs of the training scheme, and has therefore not been attempted.

In our paper, we had suggested a hybrid approach in which, at the first stage, the generative model of Reiss and Schwikowski (2004) is applied to predict *if* a protein sequence is binding, and our discriminative model is applied at the second stage to predict *which* protein the sequence binds to. Our simulations suggested that this combined scheme will outperform the generative model of Reiss and Schwikowski (2004) owing to the better performance of the proposed discriminative model at the second stage. In Chapter 4, we instead suggested an alternative discriminative model to incorporate information from the non-binding sequences.

This improved performance is presumably the consequence of two important modifications. First, the hyperparameters in our approach, which are in some way akin to the tuning parameters in the model of Reiss and Schwikowski (2004), have been integrated out. As a consequence of this integration, our scheme depends on some effective hyperparameters that are automatically updated during training (see Subsection 3.4.4). This renders our approach independent of any tweaking parameters that would otherwise have to be hand-tuned by the user. The second improvement is related to the discriminative nature of our model. Note that Reiss and Schwikowski (2004) also tries to include a discriminative feature into his generative model by penalising the detection of over-represented but non-discriminative motifs. However, this approach is rather heuristic, and introduces another user-defined tuning parameter. The model applied in our study is a proper discriminative model per se, which has been consistently derived within the probabilistic context and dispenses with the need of hand-tuning another parameter.

3.9 Future work

In this section, we further propose additional avenues of research to those that were suggested in the published paper.

3.9.1 Methodological improvements

Despite the excessive computational costs of sampling mentioned in Section 3.4.4, the problem might be still amenable to using a full sampling method such as Metropolis-Hastings or hybrid MCMC (see for instance Neal, 1996) for integrating out the uncertainty over the weights. The problem with the MAP method used is that it is known to not be invariant to reparameterisation of the parameter space. In fact, for each point in the parameter space, there always exists a one-to-one monotonic mapping of the parameters to make that point the MAP (Beal, 2003). As the log likelihood shown in Equation (3.10) separates for each individual SH3 domain, and the regularisation is done per SH3 domain, the sampling problem can be split into a series of smaller easier problem, each of which may be of a tractable size. However, sampling requires a different method to incorporate the informative initialisation from the generative model, as an ergodic Markov chain loses the memory of its initialisation. The informative initialisation could instead be kept by regularising the distance between the weights and the solution found by the generative model. Additionally, using a sampling approach could rigorously deal with changing the length of the motif using a Reversible Jump Markov Chain Monte Carlo (RJMCMC) inference scheme. RJMCMC allows changing the size of the parameter space being searched, corresponding in this case to different length motifs. See Chapter 6 where we apply a RJMCMC scheme for inference in a phylogenetic model.

In Chapter 4, the performance of the unregularised ensemble was found to be very close to that of the Laplacian regularised ensemble. The overfitting effect by the unregularised model shown in Figure 3.3 that resulted in the poorer AUROC scores in Figure 3.4 might also be removed by creating an ensemble of unregularised models. Creating an ensemble probably removes the over-fitting effect due to the bias-variance decomposition of the generalisation error – see Section 4.6 for more details. However, it is the author's opinion that a sampling scheme would be a more fruitful avenue of research, as an ensemble approach is simply an approximation to a sampling scheme.

Ferraro et al. (2006) proposed a model that jointly modelled the SH3 domain and the binding site. While their performance appears to be slightly inferior to that of the methods outlined in this thesis (see Section 4.9 for this comparison), the idea of jointly modelling the domain and the motif allows generalisation to novel SH3 domains. Incorporating this idea of building a joint model of the SH3 domain and the binding site is a promising avenue of future research, as it can directly incorporate the underlying similarity between the binding sites. See Section 4.9 for more details.

O'Flanagan et al. (2005) showed that independently modelling motif positions in DNA regulatory motifs was sub-optimal and could not fully capture the distribution over the motif instances. This could also apply to modelling the binding site motifs of the PRMs or general linear motifs in peptide sequences. Barash et al. (2003) introduced more complex models of

the motif that may better capture these dependent effects. The simplest model to implement from their proposals is a mixture of PSSM motif model, where each motif consists of a mixture of multiple normal PSSMs. Other proposals included a tree structure Bayesian network where the distribution of certain motif sites are dependent on the distribution of other motif sites. The last, most complex motif model was a combination of these, mixture of tree structured Bayesian networks.

3.9.2 More informative priors on the binding sites

The structures of some protein complexes and proteins in interactions are known (see the PDB database, H.M.Berman et al., 2000), and *a priori* it would be reasonable to expect that sites that are involved in one interaction should be more likely to be involved in other interactions. Hence, the highest prior should go to sites found in an interacting surface. This could be averaged over all the interaction partners of the protein in question. The structure of all possible protein complexes and interactions are not known however. Instead, the known structures could be used to build a classifier that predicts if a amino acid is likely to be in the interaction surface. Prediction of interaction sites using a local sequence window and a neural network was first done by Piero et al. (2002). They trained a neural network to look at a given window of the physical properties of residues and attempted to predict whether or not the central residue was part of the interaction surface with the other protein. A SVM approach was later tried by Koike and Takagi (2003) that exhibited slightly improved performance.

A second reasonable expectation is that sites on the surface of the protein are more likely to contain the motif. Again, however not all proteins have had their structure worked out. In this case, predicted structures could be used, averaged over the set of candidate structure solutions. Given that a structure, or at least a predicted structure would be known in this case, the structure of the putative motifs could also be taken directly into account. For instance, the motif could be additionally characterised by the internal bond angles between successive peptides.

Neduva et al. (2005) describe various types of peptide sequence regions that are less likely to incorporate linear binding motif such as globular domains and trans-membrane segments. This could easily be incorporated in our model by a suitable adjustment of the prior in Equation (3.6) to down-weight the *a priori* probability of such regions in the peptide sequences containing the binding site motif.

The binding sites of proteins tend to be more conserved (Nimrod et al., 2005), so giving a higher prior probability to conserved regions may help locate the binding site motifs – see Chapter 7.

3.9.3 Encoding the protein sequences as physical properties

Our discriminative model can be viewed as a Time Delay Neural Network (TDNN – see Keeler et al., 1991). This can be seen by rewriting Equation (3.9) as:

$$P(\varepsilon_{i,j}=1|s_j)=\text{logit}\left(\log\left(\frac{1}{n_j-p+1}\sum_{m=0}^{n_j-p}\exp\left\{\sum_{k=1}^{p}\left(\sum_{l=1}^{L}F_{j,l,m+k}W_{i,k,l}\right)\right\}\right)+T_i\right), \quad (3.24)$$

where $F_{j,l,n}$ is l^{th} property of the n^{th} amino acid along the j^{th} peptide sequence, while $W_{i,k,l}$ is how highly to rate the l^{th} property in the k^{th} position of the binding site motif for the i^{th} SH3 domain. L is the size of the encoding of the sequence. Equation (3.24) reduces to Equation (3.9) when a 1-of-20 encoding is used to represent the original sequence: $F_{j,l,m+k}=\mathbb{I}\left(s_{j,m+k}=l\right)$, the l^{th} property equals 1 if the l^{th} amino acid occurs in that sequence position. There are twenty amino acids, so in this encoding $L=20$.

The PRMs bind to their binding sites on the peptide sequences due to the physical properties of the binding site. Hence, representing the sequence in terms of its physical properties could improve the generalisation performance. Encoding the peptide sequence as a series of physical properties is comparatively easy as tables of the biophysical properties of the amino acid, such as the hydrophobicity, charge and surface tension, have been collated (May, 1999). For instance, $F_{j,1,n}$ could be hydrophobicity of the n^{th} amino acid in the j^{th} peptide sequence, $F_{j,2,n}$ could be the charge, etc. Due to the regularisation scheme used, these properties would probably have to be normalised in some fashion. For instance, each property could be rescaled such that it has zero mean and a variance of one when averaged across all amino acids.

Additionally, some of the properties that represent the sequence could reflect non-local characteristics of the sequence such as predictions of solvent accessibility and secondary structure. However, the predictions of the solvent accessibility and secondary structure may in turn be highly predictable given that the motif being searched for consists of a proline rich core. This is because proline is known to be highly solvent accessible, and good at breaking up secondary protein structures like α-helices and β-sheets due to its structural conformations and bond angles (Li, 2005). As these extra characteristics would be highly predictable given the proline rich core, they would not necessarily help to improve the performance of the model at distinguishing between proline rich regions. However, if the model is applied to discovering other linear motifs on sequences, these predictions may prove informative.

The binding site priors mentioned in Section 3.9.2 can be used as additional properties of the peptide sequence. This would allow the discriminative model to determine the relative importance of each of these priors, instead of needing additional parameters.

Taking more inspiration from the TDNN approach of Keeler et al. (1991), the mapping from the sequence to these properties could in turn be adaptable. This mapping can be simultaneously optimised with the weights. These properties do not have to be limited to looking at a

single amino acid position, and could instead represent some property derived from a window onto the sequence. Properties which depend on a long window of amino acids also can help model motifs that are not well modelled by a single PSSM (see end of Section 3.9.1).

3.9.4 Refining the motifs produced from other methods

The peptide motif discovering method of Neduva et al. (2005) that was described in Section 2.2.4 searched for over-represented subsequences that occur in the interactions partners of a given protein or domain. They merged together closely related binding site motifs to try and find a single consensus motif. This reduced their representation of the SH3 domain binding site motif to simply PxxP, where x represents any amino acid. The model described in this chapter focused on differentiating these closely related binding site motifs. This suggests that once the consensus binding site motif for a set of domains has been found using the method of Neduva et al. (2005), the discriminative model described in this chapter could be applied. This would serve to separate out different instances of these motifs by examining in detail the observed protein-peptide interactions related to this domain, instead of merging all interacting sequences together as done by Neduva et al. (2005).

3.9.5 A model to predict general protein-protein interactions

Deng et al. (2002) proposed a method to predict protein-protein interactions based on inferred domain-domain interactions (described in Section 2.3.2). The method described in this chapter finds motifs (or domains) based on domain-protein interactions. This suggests that these two models can be coupled together to build a general protein-protein interaction predictor, which would simultaneously learn which motifs/domains interact while detecting all motifs or domains involved in the interactions. It was found that this model suffered from serious convergence issues due to the large number of parameters in modelling the interaction between domains, and modelling the motifs/domains. The performance was found to be highly dependent on correctly pre-seeding the weights to detect known domains. The enhancements suggested in this section may significantly reduce this problem.

3.10 Chapter conclusion

The approach outlined in this chapter has shown promising performance at distinguishing between SH3 domains. However, it is computationally inefficient to apply the model to all sequences present in yeast due to its fully discriminative nature. As suggested earlier, one option is to investigate the hybrid scheme between the generative and discriminative models that was mentioned in Section 3.8. Instead, in the next chapter we propose an alternative discriminative method which can be efficiently trained on all sequences in yeast.

Symbol	Description
d_i	Represents the i^{th} SH3 domain.
s_j	Represents the j^{th} peptide sequence.
$\varepsilon_{i,j}$	Indicates if the i^{th} SH3 domain and j^{th} peptide sequence interact.
n_j	The length of the j^{th} peptide sequence.
p	The length of the motif which is searched for. We set $p = 9$.
$a_{i,j}$	The location of the binding site motif of the i^{th} SH3 domain along the j^{th} peptide sequence.
$W_{i,m,l}$	The log likelihood ratio of seeing the c^{th} amino acid in the m^{th} position of the binding site motif for the i^{th} SH3 domain, as opposed to seeing that amino acid when not in a motif.
T_i	The log likelihood ratio of a peptide sequence containing the binding site motif of the i^{th} SH3 Domain as opposed to not containing it.
$\theta_{i,k,l}$	The probability of observing the l^{th} amino acid in the k^{th} position of the binding site of the i^{th} SH3 domain.
$\theta_{0,l}$	The probability of observing the l^{th} amino acid in any position in the sequence which is not part of a motif.
l	Indexes the 20 amino acids.
i	Indexes the SH3 domains.
j	Indexes the peptide sequences.
k	Indexes motif positions. $k \in \{1,2,\ldots,p\}$.
m	Indexes possible positions one space before the start of the motif. $m \in \{0,1,\ldots,n_j - p\}$
$C_{i,j,k,l}$	$C_{i,j,k,l} = \delta(s_{j,a_{i,j}+k} = l)$. Alternatively, $C_{i,j,k,l} = 1$ if the k^{th} position of the binding motif in sequence s_j that binds to PRM domain d_i contains amino acid l. Otherwise, it is zero. From Reiss and Schwikowski (2004)

Table 3.4: Notation used in this chapter.

Chapter 4

Incorporating non-binding sequences into the detection of SH3 domain binding motifs

- An abridged version of this chapter was published as Lehrach et al. (2006b), submitted in 2005. Most of the redundant descriptions between this chapter and Chapter 3 have been removed.

4.1 Context within the thesis

In Chapter 3, we proposed a discriminative method to distinguish between the binding site motifs of different SH3 domains based on their interactions. That model was trained on all peptide sequences that were found to interact with at least a single SH3 Domain, as inclusion of all peptide sequences would have been computationally impractical. Furthermore, we proposed that a hybrid scheme using both Reiss and Schwikowski (2004) and our method that might increase performance. In this chapter we investigate an efficient and fully discriminative method that incorporates information from all non-binding sequences present – see Figure 4.1 for a comparison between the two models.

4.2 Chapter Abstract

We propose a novel discriminative method to distinguish between the binding sites of the SH3 domain. This model is compared to an alternative generative model using two different evaluation strategies. In the first evaluation strategy, non-binding sequences are excluded (as in Chapter 3). This strategy focuses on the ability of the models to distinguish between the binding sites of different SH3 domains. In the second evaluation strategy, non-binding sequences

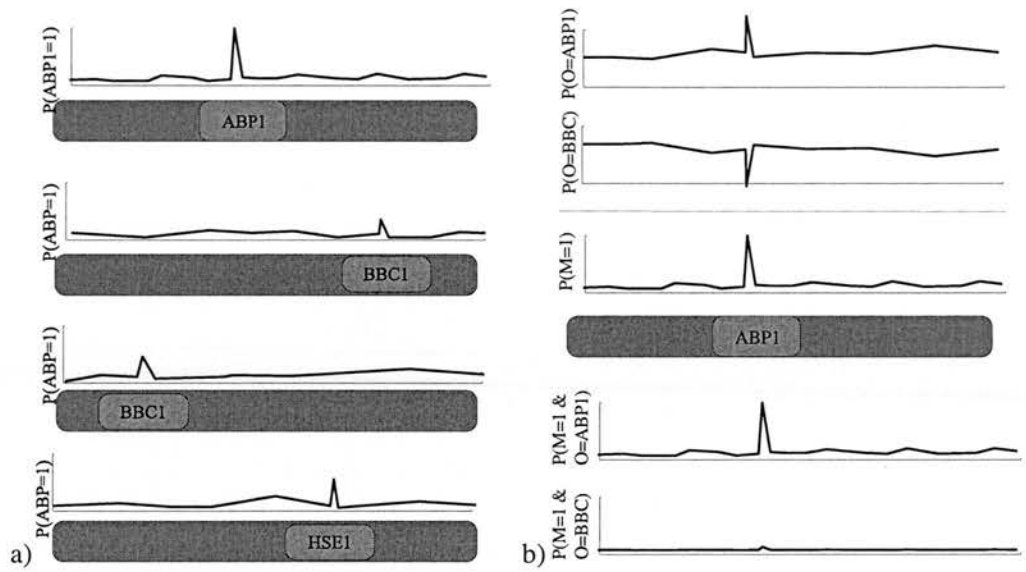


Figure 4.1: An illustration of the approach of the model described in this chapter compared to the model described in Chapter 3. Sub-figure a) demonstrates the approach taken in Chapter 3. The brown bar represents the sequence, while the blue bar represents the binding site of an SH3 domain where the label specifies which SH3 domain. Here, each graph represents a different peptide sequence. The line above the bar shows the posterior probability of the binding site of the ABP SH3 domain being detected in that position along the sequence. As all other SH3 binding site motifs are similar, they are partially detected as shown by the small resulting spikes. This decreases the incentive for the model to distinguish the binding site motif from the background. Sub-figure b) shows the approach taken in this chapter. We assume that the SH3 domains are sufficiently similar that their binding motifs can be found using a generic binding motif detector, which we represent here as $P(M = 1)$. The second assumptions is that the SH3 domain binding site only binds to a single SH3 domain, allowing us to define a distribution that ranges over the possible SH3 domains, as shown by $P(O = \text{BBC})$ and $P(O = \text{ABP1})$. The resulting predictions – $P(M = 1, O = BBC)$ and $P(M = 1, O = ABP1)$ – separately detect and discriminate between the binding site motifs, removing the small mis-detections seen earlier.

are included, and thus this strategy is more influenced by the ability of the models to detect if any SH3 domain binding sites are present. Our proposed discriminative model was found to outperform the generative model on the phage display dataset in both evaluations. On the yeast two-hybrid dataset, the generative model slightly outperformed the discriminative model when including non-binding peptide sequences. The discriminative model still outperformed the generative model when non-binding sequences where excluded, arguably showing the superior ability of the discriminative model at distinguishing between the similar binding sites motifs of the SH3 domains.

4.3 Introduction

See Section 3.3 for an overview of why modelling the binding site motifs of the SH3 domain is an interesting problem, and the potential improvements compared to the generative approach taken by Reiss and Schwikowski (2004). Our proposed model is again influenced by the protein to DNA-sequence interaction model applied by Segal et al. (2002) and Segal and Sharan (2004). However, in contrast to the model proposed in Chapter 3, the model proposed in this chapter assumes that all SH3 domains are sufficiently similar that a single binding motif detector can discover the presence of a binding site of any of the SH3 domains[1]. This consensus motif will be referred to as the generic motif. The assumption that we can detect this generic motif splits the problem into two separate stages. The first stage consists of learning to detect this generic binding site motif. Then, the second stage is learning to discriminate between the binding sites of the different SH3 domain. Detecting this generic binding motif which binds to all SH3 domains has a comparatively low computational cost. Given the presence of the binding sites, it is also then computationally inexpensive to explicitly model the differences between the binding site motifs of the various SH3 domain.

Segal et al. (2002); Segal and Sharan (2004) introduced similar discriminative approaches to modelling motifs, which they applied to DNA sequences. Due to the larger size of the alphabet (20 amino acids instead of 4 nucleotides) and the small number of interactions per SH3 domain, their maximum likelihood approach to parameter estimation is susceptible to over-fitting. An additional component of our approach, therefore, is the inclusion of a regularisation scheme, resulting in a maximum a posteriori (MAP) scheme. Additionally, we train an ensemble of models, which also reduces over-fitting.

[1] This will probably result in a loss in the discriminative power of the method, but it is computationally cheap. We propose some ways to relax this assumption in Section 4.7.

4.4 Methods

In this section we define the problem and propose a novel discriminative model to solve it – an overview of the notation used is in Table 4.3. Let $\mathbf{D} = \{d_i\}$ denote a set of SH3 domains, and $\mathbf{S} = \{s_j\}$ a set of protein sequences. We introduce a binary variable $\varepsilon_{i,j} \in \{0,1\}$, where $\varepsilon_{i,j} = 1$ indicates that sequence s_j binds to SH3 domain d_i, while $\varepsilon_{i,j} = 0$ indicates the absence of an interaction. We assume that we are given a protein interaction network $\mathbf{E} = \{\varepsilon_{i,j}, d_i \in \mathbf{D}, s_j \in \mathbf{S}\}$ from a Y2H or phage display experiment where s_j refers to the j^{th} protein sequence and $s_{j,q}$ refers to the amino acid in the q^{th} position along the j^{th} sequence. The objective is to derive a model that predicts this network from the sequences alone.

The central assumptions of our model are that the binding site motifs of the SH3 domains are sufficiently similar to allow them to be found using a single generic motif detector, that each binding site only binds to a single SH3 domain, and that sequence contains a single binding site (the OOPS model – see Section 2.2.2). These assumptions can be concisely expressed as:

$$P(\varepsilon_{i,j} = 1|s_j) = \sum_{a_j} P(M_j = 1, O_j = i, a_j|s_j), \tag{4.1}$$

where the binary variable M_j indicates if the generic motif that represents the SH3 domain binding site is present in the j^{th} sequence, $O_j \in \mathbf{D}$ represents which SH3 domain the binding site interacts with and the multinomial variable a_j represents the location of unknown position of the single binding motif on the j^{th} sequence. These assumptions make the model easy to work with but it can be seen from examining Figures 3.1 and 3.2 that some sequences violate the assumption that only a single SH3 domain binds. An obvious avenue of future research is to relax some of these assumptions.

We can split up the term on the right hand side of Equation (4.1) as follows:

$$\begin{aligned} P(O_j = i, M_j = 1, a_j = k|s_j) &= P(O_j = i|M_j = 1, a_j = k, s_j) \times \\ &\quad P(M_j = 1|a_j = k, s_j) \times \\ &\quad P(a_j = k|s_j) \end{aligned} \tag{4.2}$$

The first term in Equation (4.2) is the probability that the j^{th} sequence contains the specific binding site for the j^{th} SH3 domain in position a_j. The second term is the probability that in the j^{th} sequence, the generic binding site occurs in position a_j. The third term is the position of the binding site. a_j is *a priori* independent of the sequence. Hence, $P(a_j = k|s_j) = P(a_j = k)$. We assume that the prior over the binding motif position is uniform:

$$P(a_j = k) = \frac{1}{n_j - p + 1} \tag{4.3}$$

The first and second terms from Equation (4.2) are derived starting from generative models. We start derivation of the second term by defining the probability of the piece of sequence that

contains the generic binding site motif as:

$$P\left(s_{j,k,\ldots,k+p}|M_j=1,a_j=k\right) = \prod_{m=1}^{p}\phi_{m,s_{k+m}}, \tag{4.4}$$

where $\phi_{m,c}$ is the probability of the generic binding motif containing the c^{th} amino acid in the m^{th} position. p is the length of the motif. In order to carry out comparisons with the work of Reiss and Schwikowski (2004), we set $p=9$. We define the probability of the c^{th} amino acid occurring in any part of the sequence which is not in a motif as φ_c. Hence, the probability of a sequence which does not contain the generic binding motif is: $P(s_j|M_j=0)=\prod_{q=1}^{n_j}\varphi_{s_{j,q}}$ where n_j is the length of the j^{th} sequence. The probability of the whole sequence given that it contains the generic binding site motif at position a_j is then:

$$\begin{aligned} P(s_j|M_j=1,a_j=k) &= \overbrace{\prod_{q=1}^{k-1}\varphi_{s_{j,q}}}^{\text{Background}}\ \overbrace{\prod_{m=1}^{p}\phi_{m,s_{j,k+m}}}^{\text{Motif}}\ \overbrace{\prod_{q=k}^{n_j}\varphi_{s_{j,q}}}^{\text{Background}} && (4.5)\\ &= \prod_{m=1}^{p}\frac{\phi_{m,s_{j,k+m}}}{\varphi_{s_{j,k+m}}}\prod_{q=1}^{n_j}\varphi_{s_{j,q}} && (4.6)\\ &= B_{j,k}\prod_{q=1}^{n_j}\varphi_{s_{j,q}}, && (4.7) \end{aligned}$$

where $B_{j,k}=\prod_{m=1}^{p}\frac{\phi_{m,s_{j,k+m}}}{\varphi_{s_{j,k+m}}}$ is the likelihood ratio of the k^{th} possible motif position on the j^{th} sequence containing the generic motif as opposed to consisting only of the background distribution. Then, following the derivation in Section 3.4.2 (which in turn follows the derivation of Segal and Sharan, 2004), we can now derive the final form for the product of the second and third term of Equation (4.2). The product of the second and third term of Equation (4.2) is:

$$P(M_j=1|a_j=k,s_j)P(a_j=k)=P(M_j=1,a_j=k|s_j). \tag{4.8}$$

Applying Bayes theorem from Equation (3.8) yields:

$$\begin{aligned} P(M_j=1,a_j=k|s_j) &= \frac{P(s_j|M_j=1,a_j=k)P(M_j=1)P(a_j=k)}{\sum_{d,e}P(s_j|M_j=d,a_j=e)P(M_j=d)P(a_j=e)} \\ &= \frac{\frac{1}{n_j-p+1}B_{j,k}P(M_j=1)}{P(M_j=0)+\frac{1}{n_j-p+1}\sum_{e=1}^{n_j-p+1}B_{j,e}P(M_j=1)}. \end{aligned} \tag{4.9}$$

Dividing top and bottom by $P(M_j=1)$ yields:

$$P(M_j=1,a_j=k|s_j) = \frac{\frac{1}{n_j-p+1}B_{j,k}}{\frac{P(M_j=0)}{P(M_j=1)}+\frac{1}{n_j-p+1}\sum_{e=1}^{n_j-p+1}B_{j,e}}. \tag{4.10}$$

The first term in Equation (4.2) is the probability of a binding site binding to a specific SH3 domain. As before, we start by generatively modelling the part of the sequence which the i^{th}

SH3 domain binds to:

$$P\left(s_{j,k,\ldots,k+p}|O_j=i,a_j=k\right) \quad = \quad \prod_{m=1}^{p}\theta_{i,m,s_{k+m}}, \tag{4.11}$$

where $\theta_{i,m,c}$ is the probability of the binding site of the i^{th} SH3 domain binding site motif containing in the m^{th} position the c^{th} amino acid. This is suspect, as we are modelling the same piece of sequence with two independent variables. As this is only used in the derivation of our model we will ignore this issue here. The probability of the whole sequence is then:

$$P(s_j|O_j=i,a_j=k)=\prod_{q=1}^{n_j}\theta_{s_{j,q}}\prod_{m=1}^{p}\frac{\theta_{i,m,s_{j,k+m}}}{\varphi_{s_{j,k+m}}}. \tag{4.12}$$

Applying Bayes' Rule gives:

$$P(O_j=i|s_{j,},a_j=k)\propto P(O_j=i)\prod_{m=1}^{p}\theta_{i,m,s_{j,k+m}}. \tag{4.13}$$

For conciseness, we define $R_{j,i,k}=P(O_j=i)\prod_{m=1}^{p}\theta_{i,m,s_{j,k+m}}$. Then:

$$P(O_j=i|s_{j,},a_j=k)=\frac{R_{j,i,k}}{\sum_l R_{j,l,k}}. \tag{4.14}$$

Combining Equations (4.2), (4.10), and (4.14) gives:

$$P(O_j=i,M_j=1,a_j=k|s_j)=\frac{R_{j,i,k}}{\sum_l R_{j,l,k}}\frac{\frac{1}{n_j-p+1}B_{j,k}}{\frac{P(M_j=0)}{P(M_j=1)}+\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1}B_{j,k}}. \tag{4.15}$$

Marginalising over the unknown motif position a_j gives:

$$P(O_j=i,M_j=1|s)=\frac{\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1}\frac{R_{j,i,k}}{\sum_l R_{j,l,k}}B_{j,k}}{\frac{P(M_j=0)}{P(M_j=1)}+\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1}B_{j,k}}. \tag{4.16}$$

In order to easily train the model, it is useful to define weights which can be adjusted without worrying about constraints. We define the motif location weights $W_{m,c}=\log\frac{\theta_{m,c}}{\varphi_c}$ (where $c\in\{1,\ldots,20\}$ ranges over the possible amino acids), and the threshold as $T=\log\frac{P(M_j=1)}{P(M_j=0)}$ for detecting the generic binding motif of the SH3 domains. $B_{j,k}$ expressed in terms of $W_{m,c}$ is $B_{j,k}=\exp\left(\sum_{m=1}^{p}W_{m,s_{j,k+m}}\right)$. Using the definition of T, Equation (4.16) becomes:

$$P(O_j=i,M_j=1|s)=\frac{\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1}\frac{R_{j,i,k}}{\sum_l R_{j,l,k}}B_{j,k}}{\exp(-T)+\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1}B_{j,k}}. \tag{4.17}$$

The corresponding weights for discrimination between the SH3 domain binding sites motifs are defined as $W_{i,m,c}=\log\theta_{i,m,c}$, and their thresholds as $P(O_j=i)\propto\exp T_i$. Hence: $R_{j,i,k}=\exp\left(\sum_{m=1}^{p}W_{i,m,s_{j,k+m}}\right)\exp(T_i)$.

For $P(O_j|M_j=1,a_j,s_j)$ to be a well defined distribution, it must always hold that:

$$\sum_i P(O_j|M_j=1,a_j,s_j)=1. \tag{4.18}$$

This normalisation constraint is automatically fulfilled as $R_{j,i,k}$ always appears in the form $\frac{R_{j,i,k}}{\sum_l R_{l,i,k}}$.

4.4.1 Parameter Estimation

The parameters of the model (the set of weights $\mathbf{W} = \{W_{m,c}, W_{i,m,c}\}$ and thresholds $\mathbf{T} = \{T, T_i\}$) need to be estimated in some fashion. A standard way to optimise these parameters, adopted for instance in Segal et al. (2002) and Segal and Sharan (2004), is to follow a maximum likelihood approach. Given the training data $\mathcal{D}$, which is the set of all training sequences s_j and binding interaction indicator variables $\varepsilon_{i,j}$, we want to maximise the log likelihood:

$$\begin{aligned} -E_D &= \log P(D|\mathbf{W},\mathbf{T}) \\ &= \sum_{i,j} \quad \varepsilon_{i,j} \quad \log P(\varepsilon_{i,j}=1|s_j,\mathbf{W},\mathbf{T}) + \\ &\qquad (1-\varepsilon_{i,j}) \quad \log(1-P(\varepsilon_{i,j}=1|s_j,\mathbf{W},\mathbf{T})). \end{aligned} \tag{4.19}$$

It is straightforward to derive the partial derivatives $\frac{\partial E_\mathcal{D}}{\partial W_{m,c}}, \frac{\partial E_\mathcal{D}}{\partial W_{i,m,c}}, \frac{\partial E_\mathcal{D}}{\partial T}$ and $\frac{\partial E_\mathcal{D}}{\partial T_i}$ for Equation (4.19), which allows us to apply an iterative gradient-based optimisation scheme as done in Chapter 3. However, this direct approach ignores one of the main advantages of this model, namely that training can be performed efficiently and quickly by splitting it into separate subtasks of location and classification. First, the binding sites are located by adapting the overall weights $W_{m,c}$ and threshold T by minimising the following likelihood function:

$$\log P(D_{\text{any}}|\mathbf{W},\mathbf{T}) = \sum_j \varepsilon_j \log P(M_j=1|s_j) + (1-\varepsilon_j)\log(1-P(M_j=1|s_j)). \tag{4.20}$$

where $\varepsilon_j = 1$ if there is an interaction from any of the SH3 domains to the j^{th} sequence, and we have defined $D_{\text{any}} = \{\varepsilon_j\}$. The term:

$$P(M_j=1|s_j) = \frac{\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1} B_{j,k}}{\exp(-T) + \frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1} B_{j,k}} \tag{4.21}$$

is obtained by marginalising out O_i in Equation (4.17). Once the weights for detecting the generic SH3 domain binding site motif ($W_{m,c}$ and T) have been found, the weights to discriminate between the SH3 domains binding site motif ($W_{i,m,c}$ and T_i) are adapted using Equation (4.19). This is only performed on the peptide sequences that were experimentally found to interact with at least a single SH3 domain[2], as it is only possible to learn to discriminate between SH3 domains when at least one is predicted to bind. Consider a non-binding sequence which has not been experimentally shown to bind to a SH3 domain. However, it has a high posterior probability that a binding site motif of one of the SH3 domains is present. Training $W_{i,m,c}$ and T_i on this sequence will simply favour the solution where all SH3 domains predicted to be equally likely to bind. Hence, we only train $W_{i,m,c}$ and T_i on the sequences found to bind to at least one SH3 domain as it is an order of magnitude computationally cheaper. This is due

[2]Note that this is not all peptide sequences that interacted with at least a single SH3 domain due to our use of cross-validation to ensure we never test upon sequences which we have trained from.

to the discriminative nature of our model, and an alternative generative formulation could still learn from these sequences. Note that this does not apply when attempting to predict binding sites.

4.4.2 Approximating the joint posterior

Evaluating Equation (4.17) is computationally expensive as it is evaluated per potential motif binding position on each peptide sequence, per SH3 domain, and per optimisation step. Hence, it is evaluated at least $25 \times 10^5 \times 500$ times (there are about 10^5 potential motif positions, and we assume 500 optimisation steps). It is hence a good candidate for approximation. The intuitive idea behind the approximation is that when $B_{j,k}$ is relatively small compared to $\sum_k B_{j,k}$, then assuming that $B_{j,k} = 0$ has little effect on the result and stops us from having to compute the $\frac{R_{j,i,k}}{\sum_l R_{j,l,k}}$ terms. We introduce a binary variable $I_{j,k}$, where $I_{j,k} = 1$ indicates that the $B_{j,k}$ term remains in the approximation, and $I_{j,k}$ implies that the term is removed. Our approximation to Equation (4.17) is then:

$$\hat{P}(O_j = i, M_j = 1|s) = \frac{\frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} I_{j,k} B_{j,k}}{\exp(-T) + \frac{1}{n_j-p+1}\sum_{k=1}^{n_j-p+1} I_{j,k} B_{j,k}}. \tag{4.22}$$

Next, we specify that the approximation must still evaluate a fraction f of the mass of $B_{j,k}$ for each sequence. Hence, we choose $I_{j,k}$ such that for all j:

$$f \leq \frac{\sum_{k=1}^{n_j-p+1} I_{j,k} B_{j,k}}{\sum_{k=1}^{n_j-p+1} B_{j,k}} \tag{4.23}$$

holds, and deal with choosing f instead.

Given f, we pick $I_{j,k}$ to minimise the computational costs of the approximation. First $B_{j,k}$ is sorted, and we initialise all $I_{j,k} \leftarrow 0$. We then repeatedly set $I_{j,k} \leftarrow 1$ for the largest values of $B_{j,k}$ where $I_{j,k} = 0$ until the constraint in Equation (4.23) is satisfied. For a randomly chosen protein sequence of length 1200, we only needed two sites to capture 80% of the mass of the $B_{j,k}$, four to capture 90%, etc. We can calculate the quality of the approximation as follows. Consider the relative accuracy of the approximation:

$$\frac{\hat{P}(O_j = i, M_j = 1|s)}{P(O_j = i, M_j = 1|s)} = \left(\frac{\sum_{k=1}^{L-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} I_{j,k} B_{j,k}}{\sum_{k=1}^{L-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} B_{j,k}} \right) \left(\frac{\exp(-T) + \frac{1}{n_j-p+1}\sum_{k=1}^{L-p+1} B_{j,k}}{\exp(-T) + \frac{1}{n_j-p+1}\sum_{k=1}^{L-p+1} I_{j,k} B_{j,k}} \right). \tag{4.24}$$

Note that the first term can be upper-bounded as:

$$\sum_{k=1}^{L-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} I_{j,k} B_{j,k} \leq \sum_{k=1}^{L-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} B_{j,k}. \tag{4.25}$$

Hence:

$$\frac{\sum_{k=1}^{L-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} I_{j,k} B_{j,k}}{\sum_{k=1}^{L-p+1} \frac{R_{j,i,k}}{\sum_l R_{j,l,k}} B_{j,k}} \leq 1 \tag{4.26}$$

The second term in Equation (4.24) can be upper-bounded:

$$\frac{\exp(-T) + \frac{1}{n_j-p+1} \sum_{k=1}^{L-p+1} B_{j,k}}{\exp(-T) + \frac{1}{n_j-p+1} \sum_{k=1}^{L-p+1} I_{j,k} B_{j,k}} \leq \frac{\exp(-T) + \frac{1}{n_j-p+1} \sum_{k=1}^{L-p+1} B_{j,k}}{\exp(-T) + f \frac{1}{n_j-p+1} \sum_{k=1}^{L-p+1} B_{j,k}} \tag{4.27}$$

due to the definition of f in Equation (4.23). Our overall upper-bound is thus:

$$\frac{\hat{P}(O_j = i, M_j = 1|s)}{P(O_j = i, M_j = 1|s)} \leq \frac{\exp(-T) + \frac{1}{n_j-p+1} \sum_{k=1}^{L-p+1} B_{j,k}}{\exp(-T) + f \frac{1}{n_j-p+1} \sum_{k=1}^{L-p+1} B_{j,k}}. \tag{4.28}$$

Defining:

$$d = \frac{1}{\exp(-T)} \frac{1}{n_j - p + 1} \sum_{k=1}^{L-p+1} B_{j,k}, \tag{4.29}$$

and dividing the top and bottom of the right hand side of Equation (4.28) by $\exp(-T)$, the upper bound from Equation (4.28) becomes:

$$\frac{\hat{P}(O_j = i, M_j = 1|s)}{P(O_j = i, M_j = 1|s)} \leq \frac{1+d}{1+fd} \tag{4.30}$$

As $\lim_{d \to 0} = 1$, $\lim_{d \to \infty} = 1/f$, and $1 \leq 1/f$ (as $f \leq 1$), we redefine our upper bound as:

$$\frac{\hat{P}(O_j = i, M_j = 1|s)}{P(O_j = i, M_j = 1|s)} \leq \frac{1}{f}. \tag{4.31}$$

The relative lower bound cannot be derived in this fashion as the first term of Equation (4.24) cannot be usefully lower bounded. This can be seen by considering a situation where $R_{j,i,k} = 1 - I_{j,k}$ for $j = 1$. However, we can derive the maximum amount a probability could be underestimated. Consider:

$$\sum_i w_i + \delta = 1$$

$$\sum_i w_i = 1 - \delta, \tag{4.32}$$

$$\sum_i \hat{w}_i = 1 \tag{4.33}$$

where δ is some maximally underestimated probability, w are the other probabilities and $\hat{w}$ are the approximated probabilities. Dividing the correct and estimated probabilities, we obtain:

$$\frac{\sum_i \hat{w}}{\sum_i w} = \frac{1}{1-\delta}. \tag{4.34}$$

As we are considering the maximal underestimation, we insert the relative upper bound from Equation (4.31), giving:

$$\frac{1}{f} = \frac{1}{1-\delta}. \tag{4.35}$$

Hence, $\delta = 1 - f$. In our experiments we choose $f = 0.9$. Hence, any probability approximated using this method is over-estimated at most by a factor of 1.1, or underestimated by 0.1, whilst giving significant computational savings of around an order of magnitude. As this is the worst case analysis, we would expect that in practice the approximations will be closer to the correct answer.

4.4.3 Regularisation

The maximum likelihood approach discussed in the Section 4.4.1 suffers from a susceptibility to over-fitting. We follow the approach outlined in Section 3.4.4 to combat this over-fitting, where instead of finding the maximum likelihood solution, we instead place priors on the weights and find the Maximum A Posterior (MAP) solution.

Ten different MAP solutions are found from ten different random initialisations. The resulting models are sorted by their posterior probability and the top half are used to create an ensemble, approximating the use of a sampling scheme. A prediction for an edge is then the mean prediction from the models in the ensemble. See also Section 3.4.5 for more details of the optimisation algorithm used.

4.5 Simulations

We used the same datasets that are described in Section 3.5. Section 3.5 also describes the ROC curves, and AUROC scores that are used in this evaluation. We again refer to the model of Reiss and Schwikowski (2004) as the generative model. Their model is described in detail in Section 3.4.1.

The ability of the generative model to learn the differences between the SH3 binding domains is poorly reflected in the evaluation of Reiss and Schwikowski (2004), where all of the sequences that are not predicted to bind to an SH3 domain are included in their evaluation. As an illustration, consider a model that correctly predicts every instance of an SH3 binding domain, but without distinguishing between different SH3 domains, which corresponds to what the authors refer to as the *global* model. In terms of predicting actual protein interaction networks, as depicted for example in Figure 3.1, the performance of this predictor is poor; it will either predict an interaction with none or with all SH3 domain proteins. However, due to the large number of non-interacting sequences – Reiss and Schwikowski (2004) include about 6000 in their evaluation - its AUROC score will be close to the optimal value of 1.0. Consequently, large AUROC scores do not indicate a reliable prediction of the actual SH3 protein interaction network. To avoid this fallacy, we additionally perform a comparison where only the binding partners of the SH3 domains are included in our evaluation; these are the proteins that correspond to the nodes shown in Figures 3.1 and 3.2. This was the evaluation strategy

used in Chapter 3.

When evaluating the classifier with non-binding sequences (the same evaluation strategy used by Reiss and Schwikowski, 2004), it is of particular interest if the classifier can predict positive interactions with a low rate of false positives. In order to evaluate this, we additionally calculate the AUROC01 score, which is the proportion of the possible area filled under the ROC curve given that the rate of false positives is smaller than 0.1. Hence, when evaluating the models with non-binding sequences, the AUROC01 score is more informative. In contrast, when evaluating without non-binding sequences, the number of non-interactors is considerably reduced, and the total AUROC becomes more informative. See also Ben-Hur and Noble (2005) for a discussion of cases where AUROC is more informative versus AUROC01 or visa versa.

The models are compared against the generic SH3 domain binding motif locator, which does not distinguish between the SH3 domains but only looks for motifs indicating that any SH3 domain binding site is present. This is the classifier that results from the optimisation of Equation (4.20), where the probability of each interaction is set to be: $P(\varepsilon_{i,j}) = P(\varepsilon_i)$.

When examining the ability of the models to discriminate between SH3 domains, the naïve classifier is also shown. The naïve classifier simply predicts that the probability of each interaction occurring is the prior frequency of that SH3 domain binding to each peptide sequence:

$$P(\varepsilon_{i,j} = 1) = \frac{1}{|\mathbf{S}|} \sum_{j=1}^{|\mathbf{S}|} \varepsilon_{i,j}. \tag{4.36}$$

4.6 Results

4.6.1 Yeast two-hybrid dataset

Figure 4.2 shows the performance of the models on the yeast two-hybrid network (see Figure 3.1). Figures 4.2a and 4.2c show the performance of the models when non-binding sequences are included. In this evaluation we will take the AUROC01 score as being a better measure of the performance – see Section 4.5. Correspondingly, Figures 4.2b and 4.2d show the results when non-binding sequences are not included. Here we will take the AUROC score as being more important – again, see Section 4.5.

Whether evaluating with or without non-binding sequences, the Laplacian regularised models always outperformed the equivalent unregularised models. The simple "generic" models (see Section 4.5) were outperformed by the full models when non-binding sequences where excluded from the evaluation. This is not surprising as this evaluation focuses on the ability of the model to distinguish between the binding sites of the different SH3 domains. When evaluating with non-binding sequences, the generic models slightly outperformed the full models, as can be seen by their greater AUROC01 scores. This indicates that the full model might be over-fitting noise that is present in the dataset, or that our assumptions poorly fit the dataset.

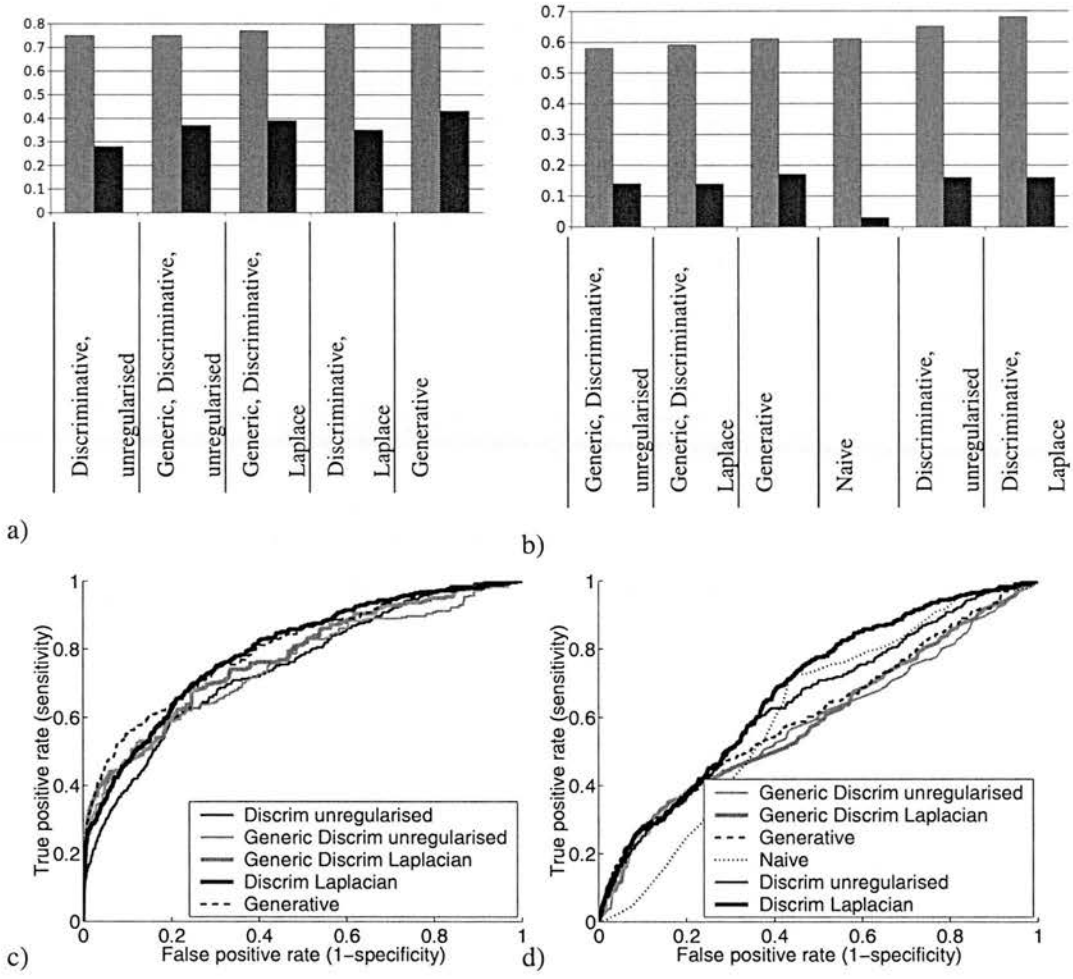


Figure 4.2: A comparison of the performance of the discriminative and generative models on the yeast two-hybrid network. Sub-figure a) shows the performance of the various classifiers when all non-binding sequences are included. The left bar for each model represents the AUROC score, while the right bar represents the AUROC01 score. Sub-figure c) shows the corresponding ROC curves. Sub-figures b) and d) show the performance of the models without non-binding sequences, which focuses the evaluation on the ability of the models to distinguish between the different SH3 domains. "Generic" refers to only using the generic SH3 domain binding site detector.

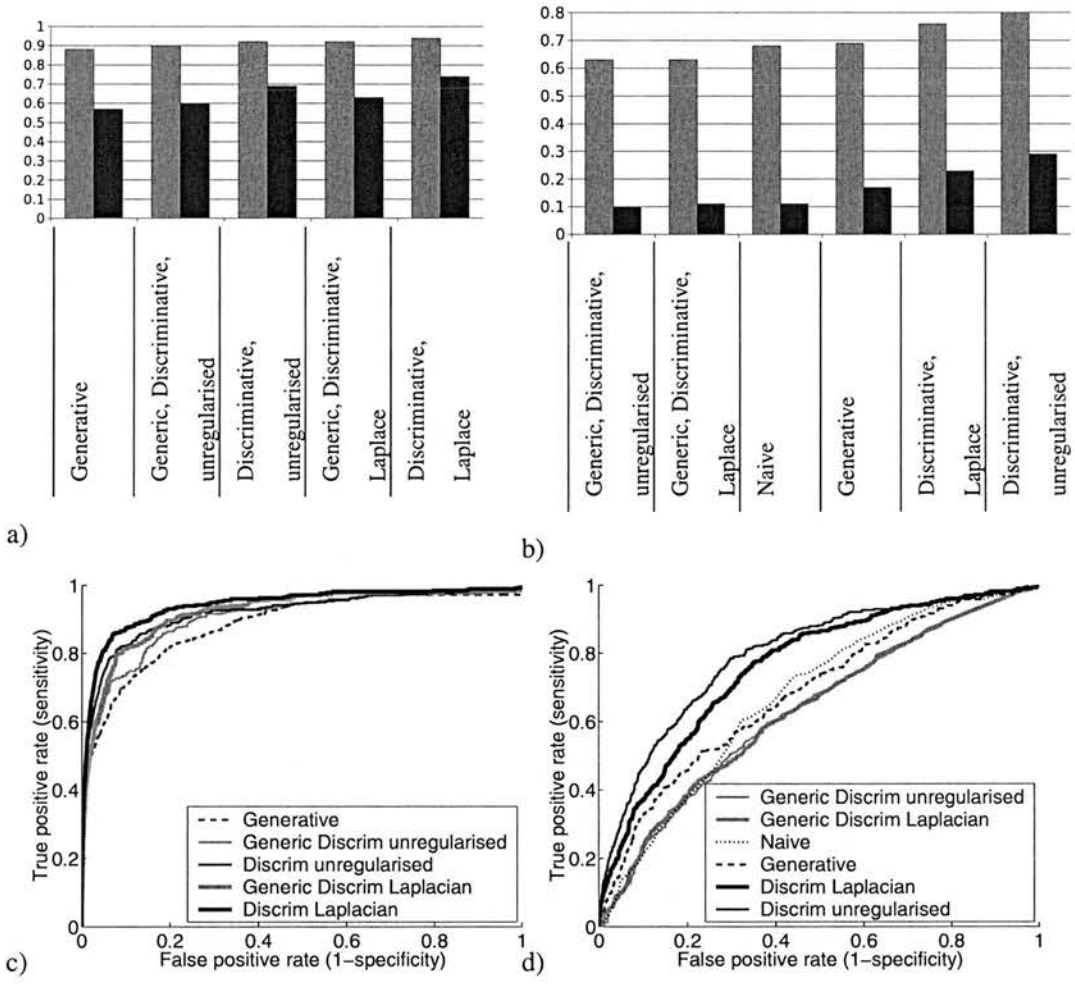


Figure 4.3: Evaluating the performance of the models on the phage display dataset, laid out in an identical fashion to Figure 4.2.

The simpler models cannot overfit the noise as they contain less parameters. This would also explain why regularisation was found to increase the performance of the models.

When evaluating with non-binding sequences, the generative model of Reiss and Schwikowski (2004) slightly outperformed the discriminative models. When the non-binding sequences are removed the full models outperformed the generative model (as measured by the AUROC score), indicating that they are probably superior at distinguishing between the binding sites motifs of the SH3 domains.

4.6.2 Phage Display dataset

Figure 4.3 summarises the results obtained on the phage display dataset of Figure 3.2. In contrast to the results on the yeast two-hybrid dataset shown in Section 4.6.1, regularisation did not always improve the performance of the model. While regularisation improved the

performance when including non-binding sequences (Figure 4.3a), it appeared to be counter-productive when discriminating between the different SH3 domains (Figure 4.3c). Note that this effect is probably a consequence of using an ensemble of models. It can be explained in terms of the bias-variance decomposition of the generalisation error, where an ensemble of individually overfitting predictors may substantially reduce the variance term; see Sollich and Krogh (1996) and Husmeier and Althoefer (1998) for further details.

The full discriminative models outperformed the simpler "generic" models. When evaluated with non-binding sequences, the AUROC01 scores of the full models are higher. When non-binding sequences are excluded, the full models always significantly outperformed the simpler "generic" models. The benefit from using the full model as opposed to only the generic binding site detector is more pronounced than on the yeast two-hybrid dataset. This is probably due to the better match between the phage display data and the model – yeast two-hybrid data is known to be very noisy.

When evaluating with non-binding sequences, all the discriminative models outperformed the generative model of Reiss and Schwikowski (2004). The simplest "generic" models slightly out-performed the generative model when the non-binding sequences where removed, but the full discriminative models significantly outperformed the generative model.

4.7 Discussion

Like Reiss and Schwikowski (2004), the present chapter has focused on the prediction of protein interactions from sequences alone. This *ab initio* approach is known to be a hard problem, as demonstrated recently by Ben-Hur and Noble (2005) and Yamanishi et al. (2005), who independently showed that a substantial improvement in the prediction performance can be achieved by integrating different types of heterogeneous post-genomic data. However, Ben-Hur and Noble (2005) also showed that non-sequence data do not distinguish between real physical interactions and the involvement of protein pairs in common pathways and complexes, while a sequence-based approach tends to detect signals that are directly related to a binding site. Consequently, improvements in *ab initio* methods, like the one investigated in our paper, are important to distinguish between co-complexed proteins and real physical interactions, and will substantially contribute to the overall problem of protein interaction prediction.

The work presented in the present chapter has been motivated by Reiss and Schwikowski (2004), who developed a probabilistic sequence model based on the Gibbs motif sampler. Our alternative discriminative approach removes the need for hand-tuning heuristic parameters, allowing easy application to novel datasets. Both models perform sufficiently well in discriminating between binding and non-binding sequences, with large AUROC scores and large slopes of the ROC curves for low false positive values. The generative model performed slightly better on the Y2H dataset when all sequences are included, while the discriminative model was better at distinguishing between the SH3 domain binding motifs for the Y2H datasets and always performed better on the phage display dataset.

The task of discriminating between the different SH3 domains is substantially harder, owing to two reasons: the training set is much smaller, and the binding motifs are quite similar (e.g., all exhibit a proline-rich core), requiring the model to pick up on subtle differences between them. The ROC curves obtained for the discriminative task (right panels in Figures 4.2 and 4.3) are noticeably better than what would have been obtained by chance. This is an encouraging finding that should stimulate future work on *in silico* prediction methods. On the discriminative task, the model proposed in this chapter outperforms the generative model on the yeast two-hybrid as well as the phage display data. Hence, it makes an important contribution towards the actual identification of the protein interaction network, as opposed to only discriminating between binding and non-binding sequences.

The model promises to provide biologically relevant information like predictions of locations of the binding sites. The discriminative basis of the model encourages focusing on distinguishing between the binding motifs that mediate the different PRM-peptide interactions, allowing the model to pick up on faint but potentially important differences which could otherwise be lost with the heuristic and parameterised discrimination used by Reiss and Schwikowski (2004).

Method	Chapter	AUROC	AUROC01
Generic only, Discriminative, unregularised	4	0.58	0.14
Generic only, Discriminative, Laplacian regularisation	4	0.59	0.14
Generative model	4	0.61	0.17
Naive	4	0.61	0.03
Discriminative, unregularised	4	0.65	0.16
Discriminative, Laplacian regularisation	4	0.68	0.16
Discriminative model, informative initialisation	3	0.67	0.17
Discriminative model, randomly initialisation	3	0.67	0.16

Table 4.1: A comparison of the performance on the yeast two-hybrid dataset between the model in Chapter 3, the model proposed in this chapter and the generative model of Reiss and Schwikowski, 2004. No none binding sequences are included in this evaluation.

Note that our model can equally be applied to the prediction of protein-DNA interactions and the identification of transcription factor binding sites. In fact, the reduced alphabet size for DNA (4 nucleotides rather than 20 amino acids) will render the over-fitting problem less severe, thereby reducing the need for the stringent regularisation scheme applied in our approach. In general, this discriminative model should be applicable to the modelling and recognition of many different regulatory elements. Other promising future work is to enhance the detection of the generic motif. This can be done by modelling it using a mixture of motifs, capturing more of variation that occurs when modelling the generic binding site. Most suggestions from Section 3.9 can also be applied to this model.

4.8 Comparison with the model in Chapter 3

Table 4.1 compares the AUROC and AUROC01 scores achieved on the yeast two-hybrid dataset between the discriminative model proposed in Chapter 3, the discriminative model proposed in this chapter and the generative model of Reiss and Schwikowski (2004). As in Chapter 3, only sequences which are predicted to bind to at least a single SH3 domain are included. Both discriminative models have a larger AUROC score than the generative model of Reiss and Schwikowski (2004), but there is no significant performance difference between the discriminative model proposed in this chapter and that proposed in the last chapter. The model proposed in this chapter not only distinguishes between SH3 domains, but is applied to all sequences present in yeast in a computationally efficient manner, unlike the discriminative model of Chapter 3.

In Table 4.2, we show the corresponding performance comparison between the models on

Method	Chapter	AUROC	AUROC01
Generic only, Discriminative unregularised	4	0.63	0.1
Generic only, Discriminative Laplacian	4	0.63	0.11
Naive	4	0.68	0.11
Generative	4	0.69	0.17
Discriminative Laplacian	4	0.76	0.23
Discriminative unregularised	4	0.8	0.29
Discriminative model, informative initialisation	3	0.83	0.44
Discriminative model, randomly initialisation	3	0.71	0.19

Table 4.2: A comparison of the on the phage display dataset between the model in Chapter 3, the model proposed in this chapter and the generative model of Reiss and Schwikowski, 2004. Again, no non-binding sequences where included in this evaluation.

the phage-display dataset, again excluding non-binding sequences. We find that the discriminative model proposed in this chapter outperforms the ensemble model from Chapter 3, but is in turn out-performed by the informatively initialised model from Chapter 3. However, the model proposed in this chapter is trained on and provides predictions for the non-binding sequences, unlike the model of Chapter 3.

4.9 Relevant literature published since the submission of paper

Ferraro et al. (2006) proposed a novel method to predict interactions between SH3 domains and peptide sequences. The principal novelty of their method is that they jointly model the SH3 domain and the peptide sequences which they bind, allowing generalisation to as yet unseen SH3 domains. In particular, they look at the structures of some known SH3 domain and peptide sequence complexes to determine which amino acids on the SH3 domain and the peptide sequence interact for the different classes of SH3 domains. An amino acid on the SH3 domain and the peptide sequence were said to be in a pairwise contact if they were sufficiently close to each other in the structure. Each of these pairwise contacts was then encoded as vector describing the physical properties of the two amino acids involved. The interaction between a putative sequence and a SH3 domain can then be described as the vector concatenation of all vectors involved in individual pairwise contacts. A neural network is trained on the resulting vector to predict if an interaction occurs between that piece of peptide sequence and the SH3 domain. Only possible candidate sites containing the consensus sequence for the motif were scanned in this fashion.

Ferraro et al. (2006) compare their method to our discriminative model outlined in Chap-

ter 3, and to the generative model of Reiss and Schwikowski (2004) on the phage display dataset. However, the AUROC value they quote for Reiss and Schwikowski (2004) is 0.79, which is the overall AUROC value found on the yeast two-hybrid method, not on the phage display method. Additionally, they do not remove the set of non-binding sequences as done in Chapter 3. Hence their comparison to Lehrach et al., 2006a also appears to be invalid. The AUROC score of 0.83 shown by Ferraro et al. (2006) should be compared to the scores in Figure 4.3, namely the generative AUROC score of 0.88, and our overall score on the phage display dataset of 0.95. This suggests that their method is performing significantly worse than the method outlined in this chapter. This may be due to their simplistic initial scanning step which identifies putative SH3 domain binding sites using regular expressions.

4.10 Chapter conclusion

The fact that the model proposed in this chapter was outperformed by the informatively initialised model from Chapter 3 suggests that a promising avenue of future research is to adapt the informative initialisation that was used for the model in Chapter 3 for the model proposed in this chapter. However, there is no longer a direct correspondence between the weights in the generative and discriminative models, making the mappings of the weights more complex. While a heuristic could be designed for mapping these weights from the generative model across to the discriminative model, in the authors opinion the phylogenetic context of each SH3 domain and binding sequences provides significant amounts of information about the location of these binding sites. Hence, the focus will now be on investigating the phylogenetic context for the peptides sequences.

Symbol	Description
d_i	Represents the i^{th} SH3 domain.
s_j	Represents the j^{th} peptide sequence.
$\varepsilon_{i,j}$	Indicates if the i^{th} SH3 domain and j^{th} peptide sequence were found to interact.
ε_j	Indicates if any SH3 domain was found to interact with the j^{th} peptide sequence.
n_j	The length of the j^{th} peptide sequence.
p	The length of the motif which is searched for. We set $p = 9$.
M_j	The variable indicate if the j^{th} sequence contains the generic binding site motif, the presence of which indicates that it binds to one of the SH3 domains.
O_j	Indicates which SH3 domain the j^{th} peptide sequence was found to bind to.
a_j	The location of the binding site motif along the j^{th} peptide sequence.
$W_{m,l}$	The log likelihood ratio of seeing the c^{th} amino acid in the m^{th} position of the generic bind site motif, as opposed to elsewhere in the sequence.
T	The log likelihood ratio of a peptide sequence containing a generic binding motif.
$W_{i,m,l}$	The log likelihood ratio of seeing the c^{th} amino acid in the m^{th} position of the binding site motif for the i^{th} SH3 domains, as opposed to elsewhere in the sequence.
T_i	The relative log likelihood of each peptide sequence containing the binding site motif of the i^{th} SH3 Domain.
c	Indexes the 20 amino acids.
i	Indexes the SH3 domains.
j	Indexes the peptide sequences.
k	Indexes the potential starting positions of the motifs. $k \in \{1,2,\ldots n_j - p + 1\}$.
m	Indexes positions in the motif. $m \in \{1,2,\ldots,p\}$.
q	Indexes the positions in a given peptide sequence. $q \in \{1,2,\ldots,n_j\}$.

Table 4.3: Notation used in this chapter.

Part II

Predicting rate variation

Chapter 5

Concepts within phylogenetics and comparative genomics

5.1 Context within thesis

Detecting the binding sites of Peptide Recognition Modules (like the SH3 domains) is a non-trivial problem due to the short length and degenerate nature of the binding site motif. However, it is known that functionally important regions of sequences, such as binding sites, tend to have a different rate of observed mutations (Nimrod et al., 2005) than non-functional regions. Hence, we will develop a method for characterising the rate along a sequence. In this chapter, we start by introducing some basic concepts in phylogenetics and comparative genomics – see Section 1.2 for other additional motivations for investigating such methods.

5.2 Introduction

Species diverge over time as illustrated in the phylogenetic tree shown in Figure 5.1a, where a phylogenetic tree is a tree that shows the evolutionary relationships within a set of species (this tree is also known as a topology). However, as we cannot look back in time, we do not know how long ago the species diverged. Instead, we have to introduce some measure of the divergence between species, and postulate the existence of unknown ancestor species that correspond to the branching points in the tree. The branch lengths in the tree then represent the amount of divergence between that pair of species instead of the time since the species split. We can then find the phylogenetic tree and set of ancestors that best explain the observed divergences between the species.

A suitable measure of divergence could be the difference in a phenotypical trait like ratios between bone lengths of adults of the different species. We can propose a phylogenetic tree and corresponding set of unknown ancestor species (with postulated bone ratios) that explain

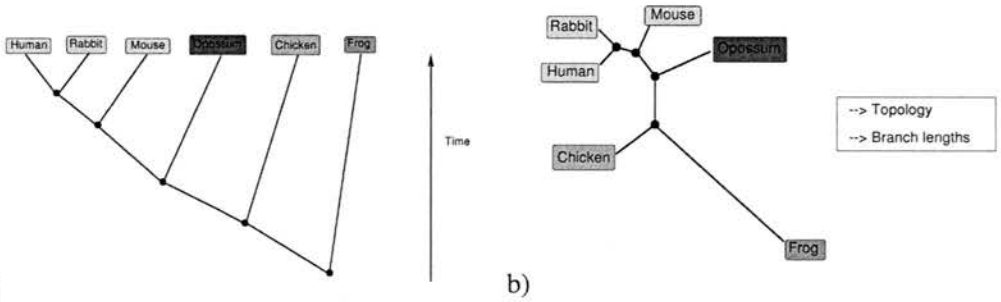


Figure 5.1: Rooted compared to unrooted phylogenetic trees. The *rooted* phylogenetic tree shown in sub-figure a) illustrates that as time progresses, a given species will diversify. Sub-figure b) shows the *unrooted* phylogenetic tree that can be inferred from the remaining species. Instead of time, the branch lengths now represent the amount of divergence between the species, and their hypothetical ancestors. Reproduced with permission from Husmeier et al. (2005).

the observed divergences between the species. The length of each branch would then correspond to the overall change in these ratios, where longer branches imply more changes. The phylogenetic tree with the shortest sum over the branch lengths would be the simplest model that explains the data and thus the most likely. This is a version of Occam's razor, which states that if two models explain the data equally well then the simpler model is the more likely explanation.

Using such phenotypical differences (like how similar the different species appear) can be misleading due to convergent evolution where different species evolve to fill the same niche. Instead, comparisons are now performed on genotypical differences, where the branch lengths in the phylogenetic tree now reflect the frequency of mutations between the underlying genomes (DNA sequences) of the species. This also gets away from the idea of the species being the only unit of evolution, as subsequences in the genome like genes can evolve in different ways to the species as a whole. For instance, bacteria can directly pass genes that confer antibiotic resistance to each other and genes can duplicate and diverge in function. This gene will then have a different evolutionary history to the organism as a whole, and the phylogenetic tree of this gene will differ from the phylogenetic tree of the corresponding species.

Naive measures of the sequence divergence, such as the number of non-matching sequence positions, are symmetrical and thus time independent. It is then impossible to determine which of the ancestral sequences are the root of the tree, as all possible choices of the root result in equally likely phylogenetic trees. Hence, unrooted trees, as shown in Figure 5.1b, are used instead.

We will focus on the simpler case of unrooted topologies in this thesis, as it is possible to recover a rooted tree from an unrooted tree by including a distant relative in the set of species.

The root is then the postulated ancestral sequence which connects to distant relative. There are also methods that can have divergence measures which are not invariant to the direction of time. This breaks the symmetry of any ancestral sequence being the root of the tree, allowing recovery of the original rooted tree. We will not focus on such methods here. See for instance Galtier and Gouy (1998) who devised a time dependent model of nucleotide evolution, and Galtier et al. (1999) where it was applied to show the common ancestor of life was not-hyperthermophilic as was thought until then.

As well as individual nucleotides along the sequences mutating, the sequences can also undergo nucleotide insertations and deletions, obscuring which nucleotides in each of the sequence correspond to each other. However, finding the best phylogenetic tree for a set of sequences requires knowing which nucleotides correspond between the sequences. An alignment of the sequences shows which nucleotides correspond between the various sequences, and is demonstrated in Figure 5.2. The best-known alignment method is ClustalW (Thompson et al., 1994). See also for instance Lassmann and Sonnhammer (2005) for an example of a recent, high performance sequence alignment method. In this thesis, we will assume that we are provided with an alignment of the sequences of interest.

Maximum Parsimony was an early method for constructing phylogenetic trees that attempted to find the best phylogenetic tree by minimising the number of mutations encountered along each branch of the tree. The intuition behind this is that models which require more mutations to explain the same alignment are more complex and thus should be penalised against. Again, this is a version of Occam's razor, which states that if two models explain the data equally well, the simpler model is the more likely explanation. While originally the lack of a model was regarded as an advantage (see Felsenstein, 2001 for an entertaining history), it was later realised that maximum parsimony is the limiting case of a simple probabilistic model. Even worse, it was found that maximum parsimony will consistently return incorrect answers in certain situations, even in the limit of infinite amounts of data being provided. The underlying reason for its failure is its implicit assumption that all branch lengths are of equal length. When the branch lengths are sufficiently unequal, the topology which minimises the number of mutations along the branches is not the correct topology, as shown in Figure 5.3. More sophisticated models will score the correct topology more highly than the topology which minimises the number of mutations, and thus will not encounter this problem.

An alternative approach is to use a distance-based method, where the sequences are summarised as a matrix of the evolutionary distance between each pair of species. There are a variety of methods for building a phylogenetic tree out of these distances. For instance, a hierarchical clustering algorithm can then be used to produce a phylogenetic tree by successively pairing together similar sequences (and then clusters). Again, the underlying cause of its failure is its assumption that all branch lengths are of approximately equal lengths, and it will again in

```
Human      GCTTGACTTCTGAGGTT
Rabbit     GCGTAACTTCACATGAT
Mouse      GCGTCATACTTGAGACGCT
Opossum    GCGTCATACTTGAGACGCT
Chicken    GCGTCATACTTGACAGGCT
Frog       GCGTCATACTACATGAGACGCT
```

a)

```
Human   GCTT--GACT---TCTGAGGTT
Rabbit  GCGT--AACT---TCACATGAT
Mouse   GCGTCATACT---TGAGACGCT
Opossum GCGTCATACT---TGAGACGCT
Chicken GCGTCATACT---TGACAGGCT
Frog    GCGTCATACTACATGAGACGCT
        ****  *****   ********
```

b)

Figure 5.2: An example DNA sequence alignment. Sub-figure a) shows an example set of unaligned sequence, while sub-figure b) shows an example inferred alignment. Notice that while nucleotides can be inserted, deleted or simply mutate in place, the alignment shows which nucleotides correspond between the different sequences. The asterisks at the bottom of the alignment mark columns where all sequences have a corresponding nucleotide. The methods in this thesis will only investigate these starred columns, where all sequences have a corresponding nucleotide.

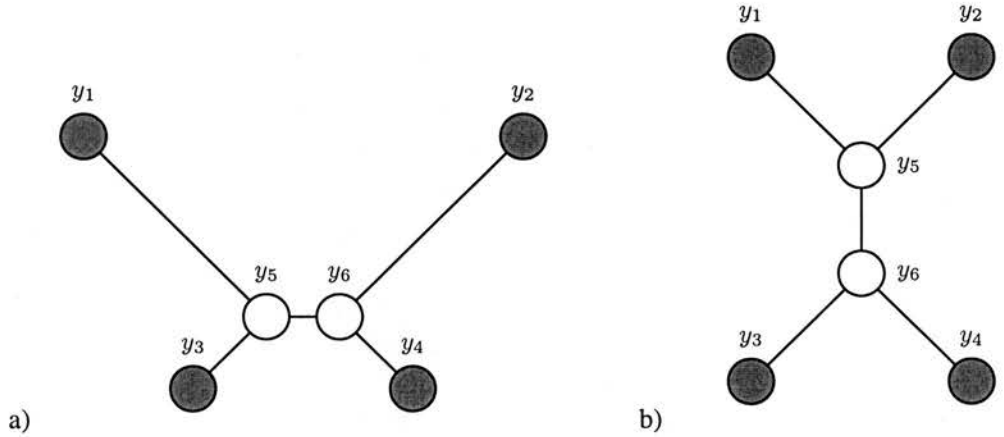


Figure 5.3: A demonstration of the short-comings of parsimony. $y_1, \ldots, y_4$ are the observed alignment positions, while y_5 and y_6 are the inferred ancestral sequences. In sub-figure a), we show the true topology. In sub-figure b), we show the best topology according to parsimony, which captures the fact that y_1 and y_2 are often the same nucleotide (there is a short path between them) at the expense of predicting the true topology. When $y_1 = y_2$, then there is still a one in four chance that $y_5 = y_6$, whereupon the figure on the right requires only a single mutation (as $y_1 = y_2 = y_5$ and $y_3 = y_4 = y_6$) while the true topology requires two mutations. Given a sufficiently unbalanced tree, parsimony will thus lead to incorrect predictions of the topology.

certain cases give the highest score to an incorrect topology. See Husmeier et al. (2005) for a more in-depth discussion of how these methods fail.

In this chapter we will focus on probabilistic methods, starting with a model of how nucleotides mutate over time. The advantage of probabilistic methods is that they are rigorous, and if our model matches the real world, are guaranteed to give the correct answer given enough data. Additionally, a model of how nucleotides mutate over time also allows us to attempt to map between the observed mutations and the amount of time that has passed since the species diverged.

5.3 Continuous time Markovian models of nucleotide evolution

A popular probabilistic model of nucleotide mutations over time is a continuous time discrete space Markov model. This does not model nucleotide insertations or deletions. Let $P(S(t))$ be the distribution over the four nucleotides at time t, where $S(t) \in \{T,C,A,G\}$. Then, for any set of time points $t_1 < t_2 < t_3 < \cdots < t_n$, a first-order **Markovian** assumption implies that:

$$P(S(t_n)\,|S(t_{n-1}),S(t_{n-2}),\ldots) = P(S(t_n)\,|S(t_{n-1}))\,. \tag{5.1}$$

This means that the model has no memory from before its current state. A further assumption is that the model is **homogeneous**. This means that the mutation process does not change over time (more complex models relax this assumption, but we will not cover them here). This implies that:

$$P(S(t+s) = j|S(t) = i) = P(S(t) = j|S(t-s) = i)\,, \tag{5.2}$$

for all s,t,i,j. The actual transitions in such a homogeneous system are described by a rate matrix $\mathbf{Q}$, where its elements $\mathbf{Q}_{i,j}$ for $j \neq i$ are proportional to the probability of a transition in interval time h as h approaches 0. We now specify the probability of a transition as:

$$P(S(t+h) = j|S(t) = i) = h\mathbf{Q}_{i,j} + o(h)\,, \tag{5.3}$$

where $j \neq i$ and $o(h)$ are all terms that decreases to 0 faster than h. $\mathbf{Q}_{i,j}$ is an instantaneous transition probability and, for a sufficiently short period of time h, is proportional to how likely a transition is to occur. The $o(h)$ term captures the fact that this transition probability must incorporate higher order terms (i.e. functions of h^2 and above), as otherwise the transition probability could go over 1 for large enough h. The smaller h is, the less of an effect these high order terms have. See, for instance, Grimmett and Stirzker (1994) for more details.

We will further make the assumption that the process is **stable**, which implies that all values of $q_{i,j}$ are finite. Also, we will assume that the process is **conservative**, which holds when $\mathbf{Q}_{i,i} = -\sum_{j\neq i}\mathbf{Q}_{i,j}$. Then:

$$P(S(t+h=i)\,|S(t) = i) = 1 - h\sum_{j\neq i}\mathbf{Q}_{i,j} + o(h)\,. \tag{5.4}$$

5.3.1 Deriving the transition matrix

Let us define $\mathbf{T}(t)$ as the transition matrix for a given time t:

$$\mathbf{T}(t) = \begin{bmatrix} P(S(t)=A|S(0)=A) & \cdots & P(S(t)=A|S(0)=T) \\ P(S(t)=G|S(0)=A) & \cdots & P(S(t)=G|S(0)=T) \\ P(S(t)=C|S(0)=A) & \cdots & P(S(t)=C|S(0)=T) \\ P(S(t)=T|S(0)=A) & \cdots & P(S(t)=T|S(0)=T) \end{bmatrix}, \tag{5.5}$$

so that Equations (5.3) and (5.4) can be stated as:

$$\mathbf{T}(t+h) = \mathbf{T}(t) + \mathbf{Q}h + o(h). \tag{5.6}$$

Set $t=0$, and we find that $\mathbf{T}(h) = \mathbf{T}(0) + \mathbf{Q}h + o(h)$, where it is obvious that $\mathbf{T}(0) = I$. Hence,

$$\mathbf{T}(h) = I + \mathbf{Q}h + o(h). \tag{5.7}$$

The Chapman-Kolomogrov equations state that the transition matrices of a homogeneous Markov process satisfy $\mathbf{T}(t+h) = \mathbf{T}(t)\mathbf{T}(h) = \mathbf{T}(h)\mathbf{T}(t)$. These equations are equivalent to simply marginalising out the nuisance variables that represent the distribution of the nucleotides at time t or h, depending on which equality is used.

Writing out the Chapman-Kolomogrov equations and substituting in Equation (5.7) gives:

$$\begin{aligned} \mathbf{T}(t+h) &= \mathbf{T}(h)\mathbf{T}(t). \\ &= (I + \mathbf{Q}h + o(h))\mathbf{T}(t) \\ &= \mathbf{T}(t) + \mathbf{Q}h\mathbf{T}(t) + o(h)\mathbf{T}(t) \\ \mathbf{T}(t+h) - \mathbf{T}(t) &= h\mathbf{Q}\mathbf{T}(t) + o(h)\mathbf{T}(t) \\ \frac{\mathbf{T}(t+h) - \mathbf{T}(t)}{h} &= \mathbf{Q}\mathbf{T}(t) + \frac{o(h)\mathbf{T}(t)}{h}. \end{aligned} \tag{5.8}$$

Taking $\lim_{h\to 0}$ of both sides gives:

$$\frac{d\mathbf{T}(t)}{dt} = \mathbf{Q}\mathbf{T}(t). \tag{5.9}$$

Further differentiating this function shows that: $\frac{d^n\mathbf{Q}(t)}{dt^n} = \mathbf{Q}^n T(t)$. The Taylor expansion of a function at position x based on its derivatives at position a is:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \cdots. \tag{5.10}$$

We know that $\mathbf{T}(0) = I$, as given that no time has passed, no mutation could have occurred. Expanding $\mathbf{T}$ around $a=0$, we find that:

$$T(t) = \sum_{i=0}^{\infty} \frac{(t\mathbf{Q})^i}{i!} = I + t\mathbf{Q} + \frac{(t\mathbf{Q})^2}{2!} + \frac{(t\mathbf{Q})^3}{3!} + \cdots. \tag{5.11}$$

This is the matrix exponential function applied to $t\mathbf{Q}$. Hence:

$$\mathbf{T}(t) = \exp(t\mathbf{Q}). \tag{5.12}$$

Substituting Equation (5.12) back into Equation (5.9) shows that it satisfies the required condition. Given a starting distribution over the nucleotides $\mathbf{x}$, the distribution over the nucleotides at time t is then $\exp(t\mathbf{Q})\mathbf{x}$. This derivation was inspired by that of Husmeier et al. (2005).

The **stationary distribution** $\pi = [\pi_T, \pi_C, \pi_C, \pi_G]^T$ is the final distribution over nucleotides that the model converges to: $\lim_{t\to\infty} \exp(t\mathbf{Q})\mathbf{x} = \pi$, where $\mathbf{x}$ is any starting distribution over the nucleotides. Notice that π does not depend on the starting distribution – this solution exists if the Markov process is ergodic (visits all states), and aperiodic (never repeats itself). The stationary distribution can be derived from its property that $\mathbf{Q}\pi = 0$.

We will concentrate on models of nucleotide evolution that are **reversible**. This implies that:

$$\begin{aligned} P(S(t+h) = i | S(t) = j) &= P(S(t) = j | S(t+h) = i) \\ &= \frac{P(S(t+h) = i | S(t) = j) P(S(t))}{P(S(t+h))} \end{aligned} \tag{5.13}$$

$$P(S(t+h)) P(S(t+h) = i | S(t) = j) = P(S(t+h) = i | S(t) = j) P(S(t)). \tag{5.14}$$

A reversible model obeys the balance equations: $\pi^T \mathbf{Q} = \mathbf{Q}\pi$. This implies that we cannot infer rooted phylogenetic trees as shown in Figure 5.1a as the model is a symmetric measure of distance. Instead, we will use such models to infer the unrooted trees shown in Figure 5.1b.

5.3.2 Normalising the branch lengths

In this project, we will normalise these rate matrices to ensure that the time parameter t represents the expected number of mutations. This gives an intuitive meaning to branch lengths. Instead of referring to the time t, we will simply refer to the weight w, where w is now the expected number of mutations. We will not look into how to calibrate the number of mutations per year to the number of mutations. The aim is to rescale the rate matrices so that the "time" t is the expected average number of mutations. We want to rescale $\mathbf{Q}$ such that the average time before transition to another state is 1, weighed by how likely it is we end up in that state. The probability of remaining in state i is:

$$P(S(t+h=i) | S(t) = i) = 1 - hq_i + o(h), \tag{5.15}$$

where we have introduced $q_i = -\mathbf{Q}_{i,i} = \sum_{j\neq i} \mathbf{Q}_{i,j}$ (otherwise identical to Equation (5.4)). Let us assume that the system is stopped after a single transition – we simply want to model the time between transitions. Then, for sufficiently small h:

$$P(S(r) = i \forall r \in (t, t+h] | S(t) = i) = P(S(t+h) = i | S(t) = i), \tag{5.16}$$

and:

$$P(S(r) = i \forall r \in (0,t+h] | S(0) = i|) =$$
$$P(S(r) = i \forall r \in (t,t+h] | S(t) = i) P(S(r) = i \forall r \in (0,t] | S(0) = i|). \quad (5.17)$$

Let $M_i(t) = P(S(r) = i \forall r \in [0,t+h] | S(0) = i)$, the probability that no mutation has occurred until time t. Then:

$$M_i(t+h) = P(S(r) = i \forall r \in (t,t+h] | S(t) = i) M_i(t) \quad (5.18)$$
$$= (1 - hq_i + o(h)) M_i(t) \quad (5.19)$$
$$M_i(t+h) = M_i(t) - hq_i M_i(t) + o(h) M_i(t) \quad (5.20)$$
$$\frac{M_i(t+h) - M_i(t)}{h} = -q_i M_i(t) + \frac{o(h) M_i(t)}{h}. \quad (5.21)$$

Take $\lim_{h \leftarrow 0}$, we find that:

$$\frac{dM_i(t)}{dt} = -q_i M_i(t).$$

We know the boundary condition $M_i(t) = 1$. Using a similar argument as in Section 5.3.1, we find that:

$$M_i(t) = \exp(-q_i t). \quad (5.22)$$

This is the probability that no mutation has occurred until time t. Hence, the probability that the mutation has occurred after time t is $1 - \exp(-q_i t)$, which is the Cumulative Density Function (CDF) of the exponential distribution with parameter q_i. The corresponding Probability Density Function (PDF) of the exponential distribution is:

$$E(t; q_i) = q_i \exp(-q_i t). \quad (5.23)$$

This is the instantaneous probability of the mutation occurring at time t, from which we can now calculate the expected time until the mutation occurs:

$$\langle t \rangle_{E(t;q_i)} = \int_0^\infty t E(t; q_i) \, dt = \int_0^\infty t q_i \exp(-q_i t) \, dt$$
$$= q_i \left[-t \frac{\exp(-q_i t)}{q_i} \right]_0^\infty - q_i \int_0^\infty \frac{\exp(-q_i t)}{-q_i} dt$$
$$= (0 - 0) - q_i \left[\frac{\exp(-q_i t)}{-q_i^2} \right]_0^\infty = \frac{1}{q_i}, \quad (5.24)$$

where we have used integration by parts ($\int f(x) g'(x) \, dx = [f(x) g(x)] - \int f'(x) g(x) \, dx$). Hence, the expected time between mutations is $\frac{1}{q_i}$, and thus the rate of mutations per unit of time starting from the i^{th} nucleotide is q_i. Recall that the amount of time spent in state i is π_i. Hence, if we want mutation to occur on average once per unit of time, then we need to ensure that the

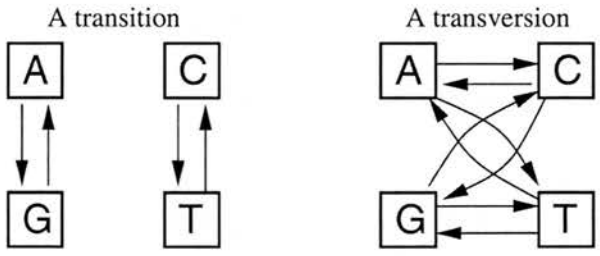


Figure 5.4: The difference between transitions and transversions.

time before a mutation occurs is on average 1, weighted by the amount of time spent in each state. This implies that: $\sum \pi_i q_i = 1$, or equivalently: $\sum_i \pi_i \mathbf{Q}_{i,i} = -1$.

In summary: given a starting distribution over the nucleotides $\mathbf{x}$, the expected distribution over the nucleotides after the time for which we have an expectation of w mutations is then $\exp\left(w\hat{\mathbf{Q}}\right)\mathbf{x}$, where $\hat{\mathbf{Q}} = \mathbf{Q}/-\sum_i \pi_i \mathbf{Q}_{i,i}$. We no longer refer to the time t, as this has been replaced by w.

5.3.3 Different models of nucleotide evolution rates

A reversible design that incorporates some biological knowledge into the rate matrix $\mathbf{Q}$ is the Kimura model, where:

$$\mathbf{Q} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} -\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha-2\beta & \beta & \beta \\ \beta & \beta & -\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha-2\beta \end{pmatrix} \end{array}. \tag{5.25}$$

This model takes into account the important differences between the rate that transitions occur and the rate that transversions occur – see Figure 5.4 for an illustration of the differences between a transition and a transversion.

However, this model assumes that all nucleotides are equally likely. This can be seen by the equilibrium distribution of $\pi = [1/4, 1/4, 1/4, 1/4]^T$. However, most DNA sequences do not contain equal numbers of the nucleotides. Hasegawa et al., 1985 remedied this problem with their HKY85 model where the rate matrix is specified as follows:

$$\mathbf{Q} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} - & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & - & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & - & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & - \end{pmatrix} \end{array}. \tag{5.26}$$

The diagonal entries are the negative sum of the other entries in that row. Notice that the equilibrium frequencies are already specified within the model, and thus can be directly set to the frequencies observed within the alignment.

Tavaré (1986) introduced the most general reversible Markovian model of nucleotide evolution where:

$$\mathbf{Q} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} - & x_1 & x_2 & x_3 \\ \pi_T x_1/\pi_C & - & x_4 & x_5 \\ \pi_T x_2/\pi_A & \pi_C x_4/\pi_A & - & x_6 \\ \pi_T x_3/\pi_G & \pi_C x_5/\pi_G & \pi_A x_6/\pi_G & - \end{pmatrix} \end{array}. \tag{5.27}$$

where again the diagonal terms are the negative sum of the other terms in the column. This rate matrix incorporates 9 parameters in total: 3 equilibrium distributions (as the equilibrium probabilities must sum to 1), and 6 rate parameters.

The rescaling required for the HKY85 model (see Section 5.3.2) is:

$$\hat{\mathbf{Q}} = \frac{1}{2\kappa(\pi_C\pi_T + \pi_A\pi_G) + 2(\pi_A + \pi_G)(\pi_C + \pi_T)}\mathbf{Q}, \tag{5.28}$$

where $\kappa = \alpha/\beta$ is the transition-transversion ratio ($\mathbf{Q}$ has been divided by β so it can be expressed in terms of κ). Hence, instead of specifying the rate of transitions α and the rate of transversions β, we now instead specify the number of expected mutations t, and the transition-transversion ratio κ. The equilibrium frequencies have to be specified in either case.

5.4 Linking nucleotide evolution to phylogenetic trees

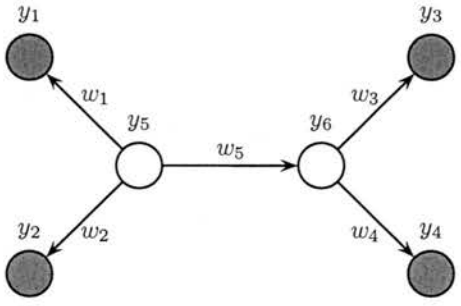
Throughout this thesis, we will assume that any nucleotides which do not occur in all of the sequences (columns containing gaps) are removed from the alignment. In Figure 5.2, the columns annotated by a star are those considered by these models.

Consider an alignment $\mathcal{D}$ of m DNA sequences, and within that consider a column where all sequences have a corresponding nucleotide. Let y_1 to y_m be the nucleotide in the m^{th} sequence. We will use the terms y_{m+1} to y_{m+a} to represent the nucleotides in the a unknown ancestral sequences corresponding to the branch points in the phylogenetic tree – see Figure 5.1). We define $\mathbf{y} = \{y_1, y_2, \ldots y_m\}$ to be the set of non-ancestral nucleotides.

The probability of column of data given the topology displayed in Figure 5.5 is:

$$P(\mathbf{y}|\psi, \mathbf{w}, \theta) = \sum_{y_5}\sum_{y_6} \pi_{y_5} \mathbf{T}(w_1)_{y_5,y_1} \mathbf{T}(w_2)_{y_5,y_2} \mathbf{T}(w_5)_{y_5,y_6} \mathbf{T}(w_3)_{y_6,y_3} \mathbf{T}(w_4)_{y_6,y_4}, \tag{5.29}$$

where $\mathbf{T}$ is defined in terms of the rate matrix $\mathbf{Q}$ in Equation (5.12), and we represent our choice of topology by ψ. Here, we define the rate matrix $\mathbf{Q}$ in terms of some evolutionary parameters θ. For instance, in the HKY85 mode, these parameters would be $\{\pi_T, \pi_G, \pi_A, \pi_C, \kappa\}$. We drop



a)

Figure 5.5: The graphical model corresponding to a 4 sequence rooted topology, with two hidden ancestral sequences. y_5 is taken to be the root sequence from which all sequences have descended. When using a reversible model of nucleotide evolution, the root sequence can be chosen arbitrarily without affecting the probability of the alignment column.

the dependency on θ to keep the notation concise. Figure 5.5 uses the notation of graphical models where filled circles represent known variables which in this case are the observed sequences. The empty circles represent the unknown ancestral sequences. The arrows indicate dependencies of each node, where arrows pointing from variable A to variable B implies that B depends on A. When an arrow points from A to B, we say that A is the parent and B is the child. This has various implications for the independencies and conditional independencies between the variables – see, for instance, Bishop, 2006 for an in-depth explanation of graphical models.

More specifically, in our model the arrows point from the copied nucleotide to the possibly mutated copy. Hence, we would never see a model where a node has multiple arrows pointing at it. As all arrows lead away from y_5, it is the root nucleotide, and all other nucleotides are mutated copies of it. The first term in Equation (5.29) specifies a distribution over y_1 given that we start from nucleotide y_5 and we expect on average w_1 mutations, while the second term specifies a distribution over y_2 given that we start from y_5 and we expect on average w_2 mutations, etc.. These terms all follows the models of nucleotide evolution discussed in Sections 5.3.1, 5.3.2 and 5.3.3.

Recall that our model of nucleotide evolution is reversible. This implies our choice of the root node is arbitrary, and the tree can be arbitrarily re-rooted without changing the probability of the alignment. We will exploit this property to simplify the evaluation of the likelihood. This also reduces the number of possible topologies, as there are $(2m-5)!!$ possible unrooted topologies (as in Figure 5.1a) for m sequences, while there are $(2m-3)!!$ possible rooted topologies (as in Figure 5.1b). $m!!$ is the double factorial of m, which is defined as follows: $m!! = m \times (m-2) \times (m-4) \times \cdots \times 1$ – see Durbin et al. (1998) for a proof. This reduction in the number of topologies for a given number of sequences also makes detecting changes in topology easier.

In general, the topology consists of a number of branches, each of which corresponds to one of the weights. We define ψ to be a function such that $\psi(1,i)$ and $\psi(2,i)$ are the start and end sequence indices of the branch that corresponds to the i^{th} weight. Hence, $y_{\psi(1,i)}$ corresponds to the nucleotide at the beginning of the i^{th} link in the tree. We can now calculate the probability of a column in the alignment:

$$P(\mathbf{y}|\psi,\mathbf{w},\theta) = \sum_{y_{m+1}} \cdots \sum_{y_{m+a}} \pi_{y_{\psi(1,i)}} \prod_{i=1}^{|\mathbf{w}|} \exp(w_k \mathbf{Q}(\theta))_{y_{\psi(1,i)},y_{\psi(2,i)}}. \tag{5.30}$$

Evaluating Equation (5.30) is computationally expensive as it requires a summation over the 4^{n-2} unknown ancestral sequences – for n sequences in the alignment, there are $n-2$ ancestral sequences. Felsenstein (1981) points out these summations can be carried out independently, as each of the summations only relate to 3 sequences.. In order to efficiently evaluate big trees, the summation signs are pushed in as much as possible. This is equivalent to ensuring that the evaluation always prunes of the tips of the tree, exposes new tips and thus proceeding inwards. For instance, Equation (5.29) would be evaluated as follows:

$$P(\mathbf{y}|\psi,\mathbf{w}) = \sum_{y_5} \pi_{y_5} \mathbf{T}(w_1)_{y_5,y_1} \mathbf{T}(w_2)_{y_5,y_2} e(y_5), \tag{5.31}$$

where the possible values of:

$$e(y_5) = \sum_{y_6} \mathbf{T}(w_5)_{y_5,y_6} \mathbf{T}(w_3)_{y_6,y_3} \mathbf{T}(w_4)_{y_6,y_4} \tag{5.32}$$

would be evaluated first. Each of these intermediate functions requires only a single summation, and can quickly be evaluated. Hence, the overall cost of evaluating the algorithm has decreased from 4^{n-2} to $4(n-2)$, a time linear with the number of hidden sequences. The saving becomes highly significant as the number of sequences increases. See Felsenstein (1981) for more details, and more examples of applying this pruning method. This method is also known as the variable elimination method (with a suitable ordering) in the machine learning field.

5.5 Maximum likelihood phylogenetic methods

We will now review methods that make two simplifying assumptions: that an alignment of the nucleotides is provided, and that all the sites in the alignment are identically and independently distributed (iid). There are flaws in this assumption due to effects like rate variation and recombination. We cover these more advanced methods in Section 5.6.

Consider again m sequences in an alignment that is N columns long. Let a column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector that contains the nucleotides at the t^{th} site of the alignment,

and $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Given a tree topology ψ, an associated vector of branch lengths $\mathbf{w}$ (or divergences along each branch in the tree), and a nucleotide substitution model (with extra parameters θ), the probability of the DNA alignment is given by:

$$P(\mathcal{D}|\psi, \mathbf{w}, \theta) = \prod_{t=1}^{N} P(\mathbf{y}_t|\psi, \mathbf{w}, \theta), \tag{5.33}$$

where $P(\mathbf{y}_t|\psi, \mathbf{w}, \theta)$ denotes the probability of the t^{th} column in the alignment – see Section 5.4 for how to calculate this. Given some model of nucleotide evolution, the task is to find the topology ψ and corresponding branch length $\mathbf{w}$ that maximise Equation 5.33:

$$\arg\max_{\mathbf{w},\psi} P(\mathcal{D}|\psi, \mathbf{w}, \theta). \tag{5.34}$$

Finding the optimal topology and branch lengths is a difficult computational problem as the number of possible topologies grows very quickly with the number of sequences. Recall that there are $(2n-5)!!$ possible unrooted topologies (as in Figure 5.1a) for n sequences, and $(2n-3)!!$ possible rooted topologies (as in Figure 5.1b).

Various heuristic methods exist that attempt to find the branch lengths and topology with the maximum likelihood. Two of the best known methods are DNAML (Felsenstein, 1981) and PUZZLE (Schmidt et al., 2002). DNAML builds a phylogenetic tree of a small subset of the sequences in the alignment, and then greedily adds on the remaining sequences. It then also attempts to remove sequences that had been added earlier and re-add them in new position in order to escape local optima. PUZZLE solves all phylogenetic trees for all selections of 4 sequences from the alignment. It then uses a heuristic to combine these mini-trees together into the overall phylogenetic tree for the full alignment.

5.6 Methods for detecting recombination and rate variation

The aim of these methods is detect recombination and/or rate variation. Recombination is a process whereby different organisms/species swap genetic material. The evolutionary history of this swapped material does not match that of the rest of the sequence. This often manifests itself as a change in topology. The methods in Section 5.6.1 attempt to detect this.

Mutation is a random process. However, along alignments of DNA sequences some regions contain more mutations. This effect is called rate variation, and occurs because mutations in functionally important regions tend to detrimentally affect the ability of the organism to pass on its genes to the next generation. Hence, mutations in these regions are less likely to fixate, and thus less likely to be conserved. We call the rate at which these mutations are conserved the conservation rate. In Section 5.6.2, we cover methods to detect these changes in the conservation rate. Finally, in Section 5.6.3 we cover methods that attempt to simultaneously detect both recombination and rate variation.

5.6.1 Detecting recombination

This is not meant to be an exhaustive overview of all possible recombination detection methods, as the main focus of the thesis is on exploiting rate variation to find motifs. See, for instance, `http://www.bioinf.manchester.ac.uk/recombination/programs.shtml` which contains an extensive list of 46 different methods that detect recombination. Our focus is on methods that detect recombination based on an alignment of the sequence, as these are the most relevant to the method used in Chapter 6.

5.6.1.1 Maximum χ^2

Maynard Smith (1992) introduced one of the earliest methods for detecting recombination. His method is based upon the χ^2 statistic, often used to test the null hypothesis that the observed frequencies match up to the expected frequencies. The χ^2 statistic is defined as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{5.35}$$

where E_i is the expected number of the i^{th} event, and O_i is the i^{th} observed number of events. The higher the value of χ^2, the bigger the discrepancy is.

The essence of the approach of Maynard Smith (1992) is to find some characteristic that varies between recombinant and non-recombinant regions. For instance, consider two parent sequences (A and B) and a recombinant (C) where the first half of C is from parent A, and the second of C is from parent B. The number of polymorphisms between recombinant C and parent A will vary significantly between the recombinant and non-recombinant regions. Thus, the method can be used to identify the breakpoint.

The characteristic is first assumed to be uniform over the whole alignment. A breakpoint is then introduced to split the alignment into two regions. The expected distributions (E_1 for the region before the breakpoint and E_2 for the region afterwards) are then calculated from the fact the characteristic (e.g. the number of polymorphic sites) is expected to have a uniform density over the sequence. O_1 and O_2 are the observed number of events before and after the breakpoint. These values are inserted in Equation (5.35), where χ^2 will now measure if the characteristic is significantly partitioned by the breakpoint, and thus that recombination has occurred. The position where χ^2 is maximised is the most likely position for recombination to have occurred. The statistical significance of finding that breakpoint is found by randomly shuffling the columns of original alignment and repeating the procedure. If the maximised χ^2 scores in the shuffles fail to match the χ^2 score found in the unshuffled data, then the finding of recombination is likely to be significant.

The main drawback with this method is that it has difficulties distinguishing between rate variation and recombination, and that the method fails on more complex recombinant structures

where more than one recombination event has occurred, resulting in multiple topology changes.

5.6.1.2 Window based methods

An alternative method to finding breakpoints is to move a window along an alignment in an attempt to find regions where the characteristics of the alignment change. This approach was used by Grassly and Holmes (1997) in their Partial Likelihoods Assessed Through Optimisation (PLATO) method. First, a consensus phylogenetic tree is built for the whole alignment. A window of columns along the alignment is then scored for how well it fits the tree. The difference between the scores in this window and rest of the alignment is measured in a window-length independent manner with the Q statistic:

$$Q_{sp} = \frac{\frac{1}{s}\sum_{t=sp}^{sp+s-1} \ln L_t}{\frac{1}{n-s}\left(\sum_{t=1}^{sp-1} \ln L_t + \sum_{t=sp+s}^{N} \ln L_t\right)}, \tag{5.36}$$

where L_t is how well the global tree fits the t^{th} column in the alignment, as calculated using Equation (5.30), and s is the length of the window. A wide range of different window lengths are tried, with s ranging from 5 to half the length of the alignment.

Areas of recombination would be expected to have a low likelihood under the consensus phylogenetic tree. This translates to low likelihood values, and thus L_t within that region will be large and negative. In the rest of the alignment, L_t will be small and negative. The negative sign of the numerator and denominator cancel, so recombinant regions will have large values of Q. In comparison, in normal regions the numerator and denominator will be approximately equal and thus cancel. Hence, Q will approximately equal 1. In order to determine statistically significant values of Q, a null hypothesis distribution of Q is simulated by shuffling the alignment, and examining the resulting distribution of over the Q values.

The primary problem with this method is that the global tree was also estimated on the recombinant regions as well as the non-recombinant region. Imagine that the recombinant region is half the sequence while the other half is the non-recombinant region. Then, Q will not vary along the alignment. The strength of this effect depends on the fraction of the alignment that is has undergone recombination. If only a small fraction of the alignment has undergone recombination, then those regions should still be detectable by this method.

McGuire et al. (1997) attempted to remedy this shortcoming by estimating the tree from the first half of the window and testing how well that tree applies to the second half of the window. However, they reduced the computational costs of the process by estimating the phylogenetic tree from the matrix of pairwise distances in the first half of the window. In this method, they pick the phylogenetic tree where the expected matrix of pairwise distances most closely resembles the actual matrix of pairwise distances observed in the alignment. This is called the

Sum of Squares (SS) statistic, and is defined as follows:

$$SS = \sum \frac{(\text{expected distances} - \text{observed distances})^2}{\text{observed distances}^2} \tag{5.37}$$

However, this method of constructing phylogenetic trees is a heuristic approach that loses information, and is known to produce incorrect results under certain circumstances. See Husmeier et al. (2005) for more details about why such heuristics for estimating phylogenetic trees fail.

First the phylogenetic tree that minimises SS for the left half of the window is found - the remaining distance is called SSa. The branch lengths of this tree are adapted to best fit the second half of the window, and the remaining distance on the second half is called SSb. The difference in squares is then $DSS^F = SSa - SSb$, where the F notes that this proceeds in a forwards direction. This is repeated with the halves of the window flipped over, where the resulting statistic is instead called DSS^B. For each window position, $DSS = \max\{DSS^F, DSS^B\}$. As with the Q statistic shown in Equation (5.36) for the PLATO method, peaks in DSS imply that the tree between the first and second half of the window differ significantly. While this can be an indication of recombination due to a topology change, it can also indicate rate variation – see McGuire and Wright (2000) for a possible method to address this problem.

Another difficulty with PLATO is that the spatial resolution of their method is poor, and locating the actual position where the topology changes is difficult as demonstrated by Husmeier et al. (2005) in their comparison of various recombination methods. It may appear that decreasing the window size decrease would help. However, as the window size decreases, it becomes harder and harder to infer sensible trees from the first half of the window, causing spurious predictions of recombination.

Husmeier and Wright (2001b) introduced a new method called the Probabilistic Divergence Measure (PDM). The principal advantages of their method compared to PLATO and TOPAL is that they rigorously deal with uncertainty in the estimation of the phylogenetic trees. The authors use sampling to infer the unknown posterior topology distribution and to marginalise out the unknown weights and parameters of the nucleotide substitution matrix.

Husmeier and Wright (2001b) use the Kullback-Leibler (KL) divergence to estimate the difference between the distribution of topologies in the window compared with the global distribution of topologies. The difference between two distributions P_i and Q_i is defined as follows:

$$KL(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i} \tag{5.38}$$

This KL score would be expected to increase in areas of recombination. Comparing the global characteristics to that found in the local windows corresponds to the methodology used in PLATO. So, in this case $P_i = P(s_P = i|\mathbf{y})$ and $Q_i = P(s_Q = i|\mathbf{y})$ where s_P is the posterior distribution of the topology within the local window, while s_Q is the posterior distribution of topology along the whole alignment.

In order to compare the distribution of topologies in the current window to the last window (as in TOPAL), they introduce a modified KL-divergence measure that took into account the distance between the compared window. Again, this modified KL score would be expected to increase in areas of recombination. The authors actually found that overlapping windows yielded the best performance.

One of the drawbacks of the PDM method is that as the number of sequences increase, the posterior distribution over the topologies becomes increasingly diffuse and uninformative. Husmeier et al. (2005) introduced a pruning step to the PDM method where uninformative topologies were removed from the posterior distribution with a post-processing clustering step. They found that this improved the performance of their model.

5.6.1.3 Hidden Markov models

The problem with the window based methods is that their spatial resolution can be limited. An alternative approach is to couple hidden Markov models with phylogenetic models. Hidden Markov models have often been used in bioinformatics – see for instance Baldi and Brunak (1998). While methods that combine phylogenetic models with hidden Markov models can often determine the location of the breakpoints with more precision (see the comparisons in Husmeier et al., 2005), the trade off tends to be that it is only possible to examine alignments containing a smaller number of sequences.

The common theme between all the methods in this section is that they represent the alignment as a hidden Markov model, where the unknown and hidden states represent the topology while the visible states represent the alignment. The hidden state that represent the topology at site t depends on the topology at site $t-1$. Hence, these models account for the fact that successive positions in the alignment are not independent and likely to share the same topology. This is also the approach taken in the model proposed in Chapter 6. Here, we will briefly review the history of such methods.

Marrying a hidden Markov model to a phylogenetic model was pioneered by Hein (1993). Their method was based on parsimony, allowing the inference to be performed in polynomial time with the length of the sequence. For the drawbacks of parismony, see Section 5.5.

McGuire et al. (2000) proposed a more rigorous probabilistic phylogenetic hidden Markov model. Let $\mathbf{y}_t$ be the t^{th} column in the alignment and s_t be the corresponding topology. The probability of the alignment given the topology at each alignment is then:

$$P(\mathbf{s}|\mathbf{y}) = P(s_1)\prod_t P(s_t|s_{t-1})\,P(\mathbf{y}_t|s_t)\,, \tag{5.39}$$

where $\mathbf{s} = \{s_1, s_2, \ldots, s_n\}$ and $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$. The task is then to find the best assignments of the topologies $\mathbf{s}$ to explain the observed alignment. Changes in the hidden topology states are then evidence that recombination has occurred. When we refer to topologies here, we do

not refer to the branch lengths, but instead only to the hierarchical set of relationships between the species.

To find the best assignment of topology states in Equation (5.39) requires a model of the emission probabilities $P(\mathbf{y}_t|s_t)$. This in turn depends on the branch lengths and parameters of the evolutionary substitution model associated with that topology. McGuire et al. (2000) estimated these per topology parameters by taking each topology in turn and fitting the branch lengths to that topology based on the whole alignment. They then used the Viterbi algorithm to find the best assignments of the topology to each alignment position, and thus topology per alignment position.

Husmeier and Wright (2001a) used the Expectation Maximisation (EM) algorithm to simultaneously optimise the weights for each topology while learning the alignment. The advantage of this approach is that the branch lengths for each individual topology were only optimised from alignments positions that actually contained that topology, in contrast to McGuire et al. (2000). Another improvement of the model of Husmeier and Wright (2001a) was that they inferred ν, the probability of the topology not changing between successive sites (see Equation (6.9) for the mapping between ν and $P(s_t|s_{t-1})$). This is useful because the frequency of recombination is not generally known in advance, and can significantly vary between different alignments.

Husmeier and Wright (2001a) is based upon EM. This has the disadvantage that parameter over-fitting can occur. This required the authors to repeat the experiment with a wide range of initial parameter setting in order to estimate of how reliable the predicted recombinations were. This test of reliability is called parametric bootstrapping.

Husmeier and Wright (2002); Husmeier and McGuire (2003) instead follow a Bayesian approach which provides full posterior distributions over the variables of interest. This also removes the need for the potentially computationally expensive parametric bootstrapping, as sampling from the posterior should be cheaper than continually having to re-optimise from random starting positions. See, for instance, Larget and Simon, 1999, who claimed that to get comparable results, bootstrapping required an order of magnitude more computational time. The Bayesian approach also has other benefits, like built-in regularisation and straightforward model comparisons. The Bayesian sampling approach taken is used in Chapter 6.

5.6.1.4 Other methods for detecting recombination

There are other methods of detecting recombination that do not depend on an alignment. For instance Boni et al. (2007) proposed a simple, fast, deterministic algorithm to determine if a sequence is a good candidate for being a recombinant of its two parents. This can be quickly applied to all possible parents and children. However, such algorithms have difficulties if the recombination has occurred a long time ago, and a significant number of mutation and splitting

events have occurred since the recombination event.

As it is easy to confuse rate variation and recombination, more sophisticated models simultaneously detect rate variation and recombination – see Section 5.6.3 for a description of such methods. We start by describing methods for detecting rate variation along alignments.

5.6.2 Incorporating rate variation

It has been found that the rate variation found amongst sites can be modelled with a gamma distribution. Yang (1993) showed how to infer the maximum likelihood phylogenetic tree in the presence of such rate variation. However, it is computationally impractical to work directly with the gamma distribution. Yang (1994) approximated the gamma distribution with a few discrete rate states, spaced such that they capture most of the variability from the gamma distribution. However, this still treated all the alignment columns as independently and identically distributed. In practice, rate variation occurs in whole regions and is thus correlated (and thus not independent) between successive columns in the alignment. Yang (1995) introduced the auto-discrete gamma method, which took these correlations into account. Felsenstein and Churchill (1996a) claimed that in attempting to approximate the auto-discrete gamma function, Yang (1995) derived from their model an auto-correlated hidden Markov model.

Felsenstein and Churchill (1996a) proposed a model that consisted of a hidden Markov model that is very similar in practice to that of Yang (1995). Their hidden Markov model selects for each column in the alignment a rate from set of predefined rate states that rescale the branch lengths of a consensus phylogenetic tree of the alignment.

It is known that the triplet of nucleotides that code for a peptide have different rates (see Section 6.5 for more details). Not taking this effect into account makes it difficult to find regional rate variation, as the HMM will only find the local site variation. Felsenstein and Churchill (1996a) suggested that the rate for a site (column in the alignment) should be the product of the rate category with a site specific offset. The authors suggest that the there should be four site-specific offset categories: three for each of the possible codon positions in an exon, and a separate offset for introns. These site-specific offsets are chosen by the user in advance.

Other research indicates that modelling rate variation might be a more complex process than just following a gamma distribution – Mayrose et al. (2005) found that the rate variation observed along alignments was better modelled as a mixture of gamma distributions, instead of by a simple gamma distribution.

It is also informative to look at how large scale identification of conserved rate regions has been carried out in practice. Margulies et al. (2007) performed a comparative genomic analysis on a large scale to identify conserved regions within part of the human genome. This involved the genomes of many different species being compared to the human genome. In order to gain

more confidence in their results and overcome shortcomings in any single individual method, Margulies et al. (2007) used four different alignment methods to generate their large alignment, and three different methods to identify areas of rate variation. We will now briefly cover the methods that were used to look for regions of rate variation.

Siepel et al. (2005) proposed a method called phastCons. This is a phylogenetic-HMM to determine rate variation along an alignment. There are two possible hidden states for each alignment position: one state indicates that the position in the alignment is neutral, while the other state indicates that the position is conserved. The amount of conservation in a conserved state (where the conservation is the shrinkage of the branch lengths compared to the neutral evolution tree) is a global parameter that is learnt using EM.

Cooper et al. (2005) proposed the Genomic Evolutionary Rate Profiling (GERP) method where every column in the alignment is independently classified as being conserved or neutral. The authors compared the observed substitution count for each site with that expected under the neutral evolution model. All alignment positions where the observed count was smaller than the expected count were classified as conserved. Neighbouring conserved position were merged into a single conserved region, where a gap of one unconstrained base would not stop conserved regions merging.

Margulies et al. (2003) introduced two different methods to detect rate variation. Both methods attempt to score the amount of conservation found with successive and overlapping 25 base windows onto the sequence. In the binomial method (BinCons), the probability of a substitution under the neutral evolution model is calculated for each species, allowing for the fact that more diverged species contribute less to this score. The cumulative binomial distribution is then used to evaluate how likely it is that the observed number of non mutated sites in the alignment would be observed under the neutral evolution model. This score is then averaged along the branches in the phylogenetic tree to account for biases in the species that are included in the alignment (e.g. multiple rodents but only a single bird). The other method introduced by Margulies et al. (2003) is based on parsimony, which was not used by Margulies et al. (2007) and so is ignored here.

5.6.3 Simultaneous detection of recombination and rate variation

Rate variation and recombination are confounding effects, so it is natural to attempt to detect both effects simultaneously, minimising the probability of confusing the effects. This section follows on closely from the hidden Markov models for detecting rate variation covered from Section 5.6.1.3.

Husmeier (2005) improved upon the model of Husmeier and Wright (2002); Husmeier and McGuire (2003) by explicitly incorporating an extra factor in the hidden Markov model to in order to capture this rate variation. See Chapter 6 for more details of this model.

Suchard et al. (2003) and Minin et al. (2005) introduced breakpoint based models for detecting recombination and rate variation. Their method introduces a series of breakpoints along the alignment, with different topologies and rates between each pair of breakpoints. We will describe their model in Section 6.6.2.

Chapter 6

A Phylogenetic Factorial Hidden Markov Model

- An abridged version of this chapter has been submitted for publication to Royal Society of Statistics, Applied Statistics journal. This work is an generalisation of an earlier published paper from Husmeier (2005) – see Section 6.4.4.

6.1 Chapter Context

We have introduced some basic concepts in phylogenetics and some methods for detecting rate variation and recombination of DNA sequences. Varying rates of conservation along an alignment can indicate where the binding site motifs of interest are. We develop a novel model for simultaneously detecting rate variation and recombination.

6.2 Chapter Abstract

The traditional approach to phylogenetic inference assumes that a single phylogenetic tree can represent the relationships and divergence amongst the taxa. However, taxa sequences exhibit varying levels of conservation, e.g. due to regulatory elements and active binding sites. Also, certain bacteria and viruses undergo interspecies recombination, where different strains exchange or transfer DNA subsequences, leading to a tree topology change. We propose a phylogenetic factorial hidden Markov model to simultaneously detect recombination and rate variation. We investigate the ability of the model to reconstruct the rate and topology of various synthetic alignments, and compare its performance to state of the art breakpoint models. Our method is applied to three DNA sequence alignments: one of maize actin genes, one bacterial (Neisseria), and the other of HIV-1. Inference is carried out in the Bayesian framework, using Reversible Jump Markov Chain Monte Carlo.

6.3 Introduction

The underlying assumption of most phylogenetic reconstruction methods is that a single phylogenetic tree captures the evolutionary history of a set of taxa. Phylogenetic trees describe the relationship among the taxa as a hierarchical tree, where the length of each branch indicates the average divergence between the associated pair of related taxa. These trees are generally estimated from an alignment of DNA or protein sequences, where the corresponding DNA or protein sequence is taken from each taxa. However in functional regions of proteins, such as the binding site for oxygen in haemoglobin or catalytically active sites in enzymes, the average divergence between the sequences decreases (Nimrod et al., 2005). Mutations in these areas are likely to adversely affect the probability of the organism surviving until reproduction, and thus these mutations are less likely to become fixed in the population. Hence, these differences in the divergence indicate areas of interest along the alignment, which is exploited in fields such as comparative genomics to find conserved regulatory elements (Chen and Blanchette, 2007). These variations are lost when the divergences between the taxa are represented by a single tree, which suggests that heterogeneity in the evolutionary rate should be explicitly taken into account. Furthermore, while the assumption of an unchanging hierarchy between the taxa is reasonable when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecies recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in the estimation of the phylogenetic tree and the rate of divergence along the sequence. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

Various methods for detecting evidence of interspecific recombination in DNA sequence alignments have been developed; see, for instance, Husmeier et al. (2005) for a recent review. The objective of the present chapter is to discuss how the performance of simultaneously detecting rate variation and recombination using a recently proposed (Husmeier, 2005) combination of phylogenetic trees with Hidden Markov models (HMMs) can be substantially improved.

HMMs provide a powerful tool widely used in Bioinformatics (Baldi and Brunak, 1998), and they have been successfully applied to the segmentation of DNA sequences (Boys et al., 2000; Boys and Henderson, 2001, 2004). Here, the objective is to locate homogeneous segments within individual DNA sequences which are compositionally different from the rest of the sequence. The hidden states represent the homogeneous segments to be detected, which are characterised by their distribution of nucleotides, or by their first-order Markovian transition probabilities between nucleotides (Boys et al., 2000). A critical question is to infer how many different segment types a DNA sequence is composed of. To this end, Boys and Hen-

derson (2001, 2004) adopted a Bayesian approach and sampled the number of hidden states from the respective posterior distribution with reversible jump (RJ) Markov chain Monte Carlo (MCMC).

The problem of detecting recombination and rate variation is related to the segmentation of a single sequence described above, but differs from it in two important aspects. First, the data to be segmented is not a single DNA sequence but a DNA sequence alignment. Second, homogeneity in a segment is not defined with respect to the nucleotide composition, but with respect to the underlying evolutionary history. This evolutionary history is captured by a phylogenetic tree, consisting of its topology H_S and associated vector of branch lengths $\mathbf{w}_{H_S}$ (depending on the nucleotide substitution model used, there might be some additional model dependent parameters θ). Hence, in generalisation of both the standard HMM applied to DNA sequence segmentations and the traditional approach to phylogenetics, one can marry the HMM to a phylogenetic tree – henceforth referred to as a phylogenetic HMM – where the latter defines the emission probabilities associated with the columns in the alignment.

Phylogenetic HMMs were originally introduced by Felsenstein and Churchill (1996b) to allow for correlations between evolutionary rates at different sites. The rates associated with the hidden states were set to *a priori* fixed values that were not inferred from the data. Siepel and Haussler (2004) applied phylogenetic HMMs to model msoaic structures in DNA sequence alignments in the context of comparative genomics. In this context, mosaic structures in an alignment imply that the alignment contains regions of a different topology or rate. The parameters were inferred by maximum likelihood in a supervised way, assuming that the hidden state sequences were known. The application of phylogenetic HMMs to the detection of recombination was first proposed by McGuire et al. (2000), with subsequent improvements of the inference methodology by Husmeier and Wright (2001a) and Husmeier and McGuire (2003). However, these models can confuse regions subject to recombination and rate variation. Husmeier (2005) addressed this problem by introducing a phylogenetic factorial HMM (FHMM), with two different types of hidden states. This disentangles topology changes – indicative of recombination – from changes of the nucleotide substitution rate. For the latter, a set of fixed, *a priori* chosen values was used, akin to the approach of Felsenstein and Churchill (1996b). This set of fixed rates limits the accuracy to which the rate variation along the sequence can be characterised.

The present chapter improves on the approach of Husmeier (2005) in three important respects. First, rather than setting the parameters associated with the hidden states to *a priori* selected fixed values, we sample them from the posterior distribution with MCMC. Second, we infer the number of hidden states, which corresponds to the number of homogeneous segments in the DNA sequence alignment, with RJMCMC. Finally, we also apply this inference scheme to allow for changes in the transition-transversion ratio along the alignment.

6.4 The Model

Our notation is summarised in Table 6.3.

6.4.1 The Bayesian phylogenetic factorial hidden Markov model (FHMM)

Consider an alignment $\mathcal{D}$ of m DNA sequences, N nucleotides long. Let a column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector that contains the nucleotides at the t^{th} site of the alignment, and $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. The traditional approach to phylogenetics assumes that sites in the DNA sequence alignment are identically and independently distributed (iid); see, for instance, Durbin et al. (1998) or Husmeier et al. (2005) for a review. Given a tree topology H_S (the notation will become clear later), an associated vector of branch lengths $\mathbf{w}$, and a nucleotide substitution model (with extra parameters θ), the probability of the DNA sequence alignment is given by:

$$P(\mathcal{D}|H_S, \mathbf{w}, \theta) = \prod_{t=1}^{N} P(\mathbf{y}_t|H_S, \mathbf{w}, \theta) \tag{6.1}$$

$P(\mathbf{y}_t|H_S, \mathbf{w}, \theta)$ denotes the probability of the t^{th} column in the alignment, and is defined by the nucleotide substitution model used. In this chapter, we use HKY85, the reversible Markov process model introduced by Hasegawa et al. (1985), which has the nucleotide substitution rate matrix $\mathbf{N}$:

$$\mathbf{N} = \begin{pmatrix} - & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & - & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & - & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & - \end{pmatrix} \tag{6.2}$$

where π_A, π_C, π_G and π_T are the equilibrium probabilities of the nucleotides, α and β are the transition and transversion rates, and the four rows and columns of the matrix refer to the four nucleotides in the order thymine (T), cytosine (C), adenine (A) and guanine (G). Each row of $\mathbf{N}$ sums to 0 – hence the diagonal "-" entries are the negative sum of the other entries in each row. The transition probability for a given time t is then $\exp(\mathbf{N}t)$, where the entries of this matrix give the probability of each nucleotide mutating into each of the other nucleotides in a given time t.

Dividing $\mathbf{N}$ by β allows $\mathbf{N}$ to be expressed in terms of $\kappa = \alpha/\beta$. Instead of using κ like Husmeier (2005) and Minin et al. (2005), we follow DNAML (Felsenstein, 1981) and PUZZLE (Schmidt et al., 2002) and use

$$E = \kappa \frac{(\pi_T\pi_C + \pi_A\pi_G)}{(\pi_T + \pi_C)(\pi_A + \pi_G)}, \tag{6.3}$$

the ratio of expected transition mutation events to transversion mutation events (Rosenberg et al., 2003), which we will refer to as the transition-transversion ratio. We define $\tau = \log_{10} E$

as it is more natural to deal with ratios in the log space. We also normalise $\mathbf{N}$ to ensure that the branch lengths $\mathbf{w}$ represent the expected number of mutations. This is done by rescaling $\mathbf{N}$ such that $\sum_{i,i} \pi_i \mathbf{N}_{i,i} = -1$ (see Section 5.3.2). Calculating the rescaling needed by referring to the appropriate terms from Equation (6.2), we see that our new $\hat{\mathbf{N}}$ is:

$$\hat{\mathbf{N}} = \frac{1}{2\kappa(\pi_C\pi_T + \pi_A\pi_G) + 2(\pi_A + \pi_G)(\pi_C + \pi_T)}\mathbf{N}. \tag{6.4}$$

To simplify the model, we follow Suchard et al. (2003) and impose a product of independent exponential distributions as a prior on the branch lengths $\mathbf{w}$:

$$P(\mathbf{w}|r) = \prod_i P(w_i|r,\theta);\ P(w_i|r) = r^{-1}\exp(-w_i/r) \tag{6.5}$$

where the w_is are the lengths of the individual branches of the phylogenetic tree. This prior is conjugate to the likelihood and makes analytical integration of the branch lengths tractable:

$$\begin{aligned} P(\mathbf{y}_t|H_S,r,\theta) &= \int P(\mathbf{y}_t|H_S,\mathbf{w},\theta)\prod_i P(w_i|r)\,d\mathbf{w} \\ &= \pi_{y_{t,1}} \sum_{y_{t,m+1}} \cdots \sum_{y_{t,m+a}} \prod_{(i,j)\in H_S} \mathbf{M}_{y_{t,i},y_{t,j}}(r) \end{aligned} \tag{6.6}$$

where $y_{t,m+1}$ to $y_{t,m+a}$ represent the nucleotides at position t of the a unknown ancestral sequences, and H_S describes the tree as a list of connections between the sequences defined by the topology of the phylogenetic tree. Suchard et al. (2003) have derived that:

$$\mathbf{M}(r) = \sum_{i=1}^{4} (1 + r\lambda_i/\beta)^{-1}\mathbf{u}_i\mathbf{v}_i^T, \tag{6.7}$$

where $\mathbf{N} = \sum_{i=1}^{4} \lambda_i \mathbf{u}_i \mathbf{v}_i^T$ is a decomposition of the nucleotide substitution matrix in Equation (6.2) and $\mathbf{u}_i$, $\mathbf{v}_i$ and λ_i are all functions of θ, the parameters of the nucleotide substitution model – see Hasegawa et al. (1985). We use the pruning algorithm of (Felsenstein, 1981) to efficiently calculate these terms in polynomial time. The hyperparameter r is a scale parameter representing the expected number of mutations over all branches. We define $\rho_R = \log_{10} r$, as an uninformative prior for a scale parameter is uniform on a log scale.

The iid assumption underlying Equation (6.1) is violated in the presence of recombination, rate variation or changes in the transition-transversion ratio. We first look at modelling topology changes caused by recombination. We assume we know the set of possible tree topologies $\boldsymbol{\rho}_S = \{\rho_{S,1}, \ldots, \rho_{S,k_S}\}$. For simplicity of implementation, we follow Husmeier (2005) and deal only with 4 sequences in the alignment, so $\boldsymbol{\rho}_S = \{\rho_{S,1}, \rho_{S,2}, \rho_{S,3}\}$ exhaustively covers all possible unrooted tree topologies. A method to select a suitable candidate set $\boldsymbol{\rho}_S$ for alignments with more sequences is suggested in Minin et al. (2005) and straightforward to integrate into our model (although our current software does not support it yet). We introduce the site-dependent discrete hidden state $H_{S,t}$, where $H_{S,t} \in \boldsymbol{\rho}_S$ represents the topology for site t. So, if $H_{S,t} = \rho_{S,i}$, then the topology at alignment position t is $\rho_{S,i}$.

Given a discrete set of the logs of the possible mean branch lengths (we also call them rates) $\boldsymbol{\rho}_R = \{\rho_{R,1}, \ldots, \rho_{R,k_R}\}$ and a set of discrete log transition-transversion ratios $\boldsymbol{\rho}_T = \{\rho_{T,1}, \ldots, \rho_{T,k_T}\}$, we allow for rate variation and changes in the transition-transversion ratio by associating each alignment position t with hidden random variables $H_{R,t} \in \boldsymbol{\rho}_R$ and $H_{T,t} \in \boldsymbol{\rho}_T$. These pick a rate ρ_R from $\boldsymbol{\rho}_R$ and a log transition-transversion ratio from $\boldsymbol{\rho}_T$ respectively. As before, if $H_{R,t} = \rho_{R,i}$ and $H_{T,t} = \rho_{T,j}$, then at alignment position t the rate is $\rho_{R,i}$ and the log transition transversion-ratio is $\rho_{T,j}$. This allows the mean rate and the transition-transversion ratio to vary along the sequence alignment. To summarise, for a site t in the alignment there are in total three hidden variables[1]: $H_{S,t}$, $H_{R,t}$ and $H_{T,t}$, giving us three *a priori* independent hidden chains.

In this chapter, the subscript A will be used to refer to any $A \in \{S,R,T\}$, where S, R and T refer to states that represent the different tree topologies, rates, and transition-transversion ratios, respectively. We use this notation to define the chains of hidden variables: $\mathbf{H}_A = \{H_{A,1}, \ldots, H_{A,N}\}$, and we represent all hidden variables in the model by defining $\mathbf{h} = \{\mathbf{H}_S, \mathbf{H}_R, \mathbf{H}_T\}$. To allow for correlations between sites that are close together in the sequence – while keeping the computational complexity limited – a Markovian dependence structure is introduced:

$$P(\mathbf{H}_A|k_A, \boldsymbol{\rho}_A) = P(H_{A,1}, \ldots, H_{A,N}|k_A, \boldsymbol{\rho}_A) = \prod_{t=2}^{N} P(H_{A,t}|H_{A,t-1}, k_A, \boldsymbol{\rho}_A) P(H_{A,1}|k_A, \boldsymbol{\rho}_A) \quad (6.8)$$

where again $A \in \{S,R,T\}$. Following Felsenstein and Churchill (1996b), the transition probabilities $\boldsymbol{\nu} = \{\nu_S, \nu_R, \nu_T\}$ are defined as:

$$P(H_{A,t}|H_{A,t-1}, \nu_A, k_A, \boldsymbol{\rho}_A) = (\nu_A)^{\mathbb{I}(H_{A,t}=H_{A,t-1})} \left(\frac{1-\nu_A}{k_A - 1}\right)^{[1-\mathbb{I}(H_{A,t}=H_{A,t-1})]} \quad \text{for } k_A > 1, \quad (6.9)$$

where $\mathbb{I}(\cdot)$ is 1 if the condition is satisfied and 0 otherwise. The parameters ν_S, ν_R and ν_T denote the probabilities of the tree topology, rate and transition-transversion ratio, respectively, not changing between adjacent sites. We follow Husmeier and Wright (2001a) and set the initial state probabilities to:

$$P(H_{A,1}|k_A) = \frac{1}{k_A} \quad (6.10)$$

The resulting model is a FHMM – as illustrated in Figure 6.1 – containing three *a priori* independent chains of hidden states, $\mathbf{H}_S$, $\mathbf{H}_R$ and $\mathbf{H}_T$, for the tree topologies, evolutionary rates and transition-transversion ratios, respectively. The probability of a column of nucleotides in the alignment, the so-called emission probability, depends on all three hidden states: $P(\mathbf{y}_t|H_{S,t}, H_{R,t}, H_{T,t})$, which can then be calculated using Equation (6.6). It also depends on the equilibrium frequencies π_A, π_C, π_G and π_T (note that the λ_i's, $\mathbf{u}_i$'s and $\mathbf{v}_i$'s in the decomposition of $\mathbf{N}$, as stated below Equation (6.6), depend on the transition-transversion ratio and the equilibrium frequencies; see Hasegawa et al. (1985)). However, we leave this dependence out to keep the notation simple. We also summarise our model in the form of a probabilistic graphical model in Figure 6.2.

[1] Note that for notational conciseness, we have merged hidden states and their associated parameters into quan-

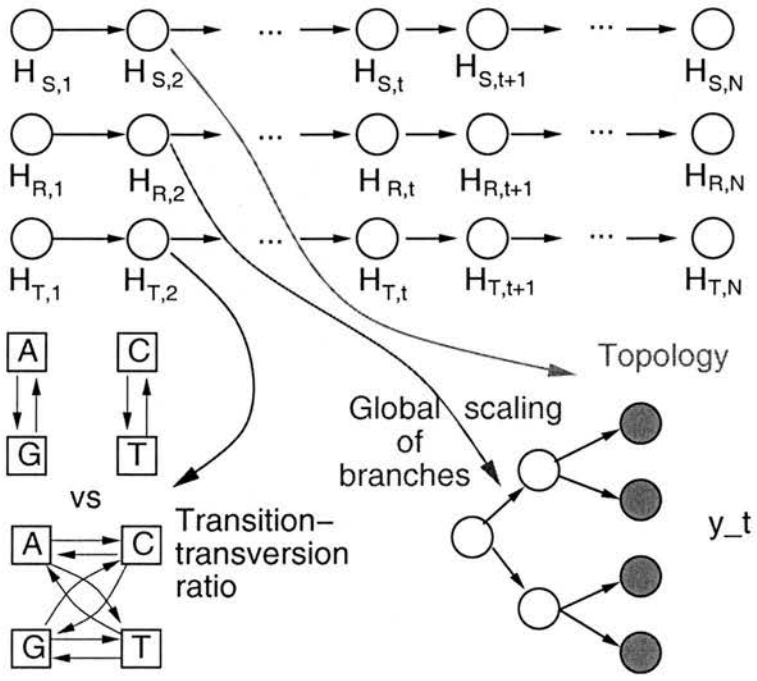


Figure 6.1: Illustration of the factorial hidden Markov nature of our model. Empty circles represent parameters or hidden variables while filled circles indicate observed variables. For each position in the alignment, there are three hidden variables representing the topology, the evolutionary rate and the transition-transversion ratio, and that these hidden variables are correlated between neighbouring positions. These specify the characteristics of the phylogenetic tree, shown in the bottom right, in which empty nodes represent the nucleotides of unobserved ancestral sequences, while shaded nodes represent nucleotides in the DNA sequence alignment. The topology $H_{S,t}$ specifies the connectivity of this tree while the log rate $H_{R,t}$ specifies how likely mutations are along each branch in the tree. The relative likelihood of seeing a transition as opposed to a transversion along each branch is specified by $H_{T,t}$ – the difference is illustrated in the bottom left (squares represent nucleotides).

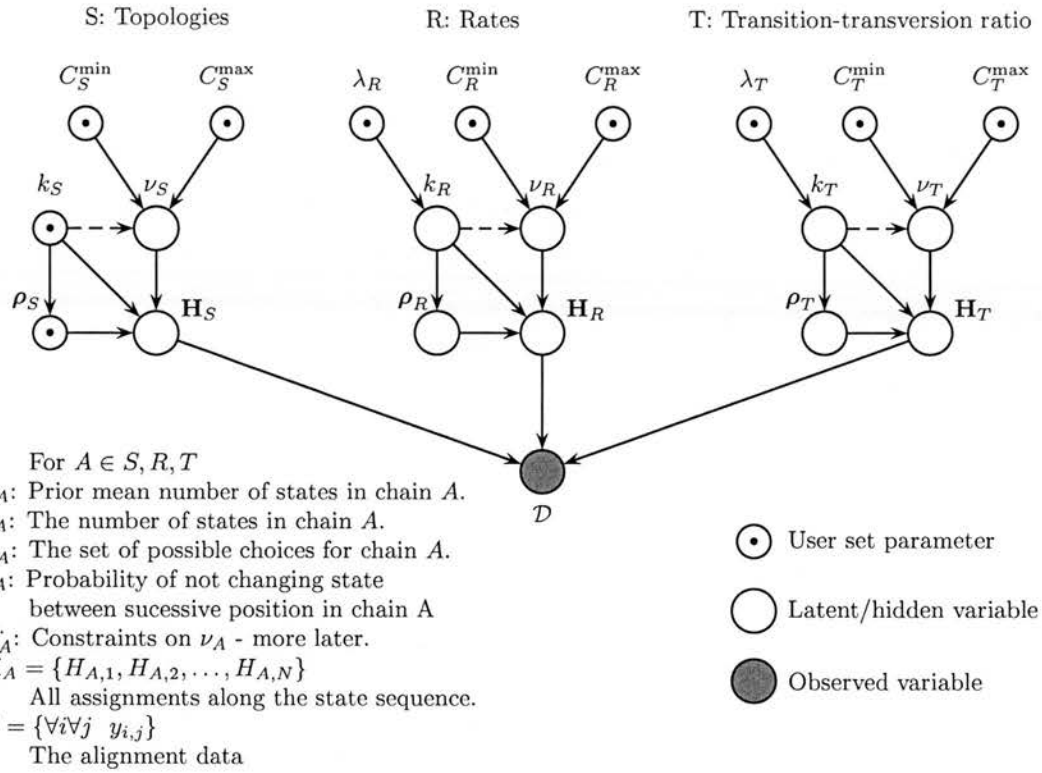


Figure 6.2: The full graphical model of the phylogenetic FHMM. Empty circles represent parameters or hidden variables, filled circles indicate observed variables, and dotted circles indicate specified parameters. We summarise our model in the form of a probabilistic graphical model (Pearl, 1988), where $\mathbf{H}_S$, $\mathbf{H}_R$, $\mathbf{H}_T$ and $\mathcal{D}$ are all chains of hidden states, as shown in Figure 6.1a. Note that k_S and $\boldsymbol{\rho}_S$ are not inferred by the construction of our model. ν_A is not defined when $k_A = 1$, which we symbolise with a dashed line.

6.4.2 Prior distributions

We introduce prior probabilities on the transition parameters $\boldsymbol{\nu}$: $P(\nu_S)$, $P(\nu_R)$ and $P(\nu_T)$. As shown in Husmeier and McGuire (2003), the conjugate prior is a beta distribution:

$$\mathcal{B}(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{6.11}$$

whose shape is determined by the hyperparameters α and β. The normalisation constant for this distribution is defined of terms of gamma functions, which are defined as follows: $\Gamma(z) = \int_0^\infty t^{z-1}e^{t-1}\mathrm{d}t$.

In the present work, we set $\alpha = \beta = 1$, reducing our prior to a uniform distribution over the interval [0,1], where we additionally constrain the range of valid values:

$$P(\nu_A) \propto \mathcal{B}(\nu_A|\alpha,\beta)\mathbb{I}\left(C_A^{\min} \leq \nu_A \leq C_A^{\max}\right), \tag{6.12}$$

the reason for this will become clear in Sections 6.9.2 and 6.9.1. We define $\mathbf{C}$ to be the set of all such thresholds: $\mathbf{C} = \left\{C_A^{\min}, C_A^{\max} | A \in \{S,R,T\}\right\}$. ν_A defines a geometric distribution over n_A, the segment length:

$$P(n_A) = (\nu_A)^{n_A-1}(1-\nu_A). \tag{6.13}$$

Hence, $\langle n_A \rangle$, the average segment length can be derived as follows:

$$\begin{aligned}\langle n_A \rangle = \sum_{n=0}^{\infty} nP(n) = (1-\nu_A)\sum_{n=0}^{\infty} n(\nu_A)^{n-1} = (1-\nu_A)\frac{d}{dv}\sum_{n=0}^{\infty}(\nu_A)^n \\ = (1-\nu_A)\frac{d}{dv}\frac{1}{1-\nu_A} = \frac{1}{1-\nu_A}\end{aligned} \tag{6.14}$$

Thus setting $C_A^{\min}$ and $C_A^{\max}$ implies that the average segment length is between $1/1-C_A^{\min}$ and $1/1-C_A^{\max}$, allowing an intuitive specification of prior knowledge. Also, the posterior distributions of ν_S, ν_R and ν_T can contain multiple modes, which can be easily selected and more closely investigated by setting $C_A^{\min}$ and $C_A^{\max}$ appropriately.

We set our prior belief on k_R, the number of rate states and k_T, the number of transition-transversion ratios to be:

$$P(k_R) \propto \frac{(\lambda_R)^{(k_R-1)}}{(k_R-1)!}\mathbb{I}(k_R - 1 \leq k_{\max}) \qquad P(k_T) \propto \frac{(\lambda_T)^{(k_T-1)}}{(k_T-1)!}\mathbb{I}(k_T - 1 \leq k_{\max}) \tag{6.15}$$

which are truncated Poisson distributions over the number of additional rate and transition-transversion ratio states. These are distributions over the number of additional states as the model is nonsensical without at least a single rate and transition-transversion ratio. We expect an average of λ_R additional rate states and λ_T additional transition-transversion states. For instance, if we expect that on average a new rate state occurs every thousand alignment columns, then we could set $\lambda_R = \frac{N}{1000}$. These priors are truncated Poisson distributions, where

tities that we refer to as "hidden variables".

the truncation ($k_{\max} = 15$ in our case) reflects our desire for a parsimonious solution, with re-use of rates and transition-transversion ratios for different parts of the alignment. In practice, we never observed k_R or k_T to be as high as $k_{\max}$, implying that our model was not affected by this truncation.

To complete the specification of our probabilistic model, we specify priors for $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$, which have k_R and k_T entries respectively:

$$P(\boldsymbol{\rho}_A|k_A) = \prod_{i=1}^{k_A} Q_A(\rho_{A,i}), \tag{6.16}$$

One choice for Q_A we investigate is $Q_A(\rho_{A,i}) = \mathcal{N}(\rho_{A,i};\mu_A,\sigma_A^2)$ where:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6.17}$$

is the probability density function of the Gaussian distribution. For comparability to Minin et al. (2005), we set these hyperparameters to the values used by Minin et al. (2005): $\mu_R = -2\log_{10} e$, $\sigma_R^2 = 2\log_{10} e$, $\mu_T = 2\log_{10} e$ and $\sigma_T^2 = 1\log_{10} e$.

Note that the likelihood of the model given the data is unchanged if we reorder the entries in $\boldsymbol{\rho}_R$ or $\boldsymbol{\rho}_T$. Hence, the posterior exhibits large degrees of symmetry. One way to break this symmetry is to use ordered but otherwise independent priors on $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$:

$$P(\boldsymbol{\rho}_A|k_A) = \mathbb{I}(\rho_{A,1} \leq \rho_{A,2} \leq \cdots \leq \rho_{A,k_A})(k_A!)\prod_{i=1}^{k_A} Q_A(\rho_{A,i}) \tag{6.18}$$

where Q_A is the distribution over a single entry, for $A \in \{R,T\}$. The $k_A!$ term compensates for the amount of space excluded by the ordering. While imposing an ordering on the prior distribution may appear intuitive, it does not affect the posterior distributions for any other variable in the model. For instance, the choice of prior does not affect the posteroir probability $P(\mathcal{D}|k_A)$– see Appendix A. Hence, we simply use the unordered prior from Equation (6.16), and note that if the posterior distribution of an individual component $\rho_{A,i}$ is of interest, post simulation relabelling procedure should be performed as described for instance in Jasra et al. (2005).

A reasonable alternative *a priori* belief about $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ is that neighbouring factors are unlikely to be very similar. We investigate using even-numbered order statistics (Green, 1995) as a prior on $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ which has the effect of penalising rate states or transition-transversion ratios that get too close to each other. This prior is also an ordered prior, and again constrains $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ such that $\rho_{R,1} < \rho_{R,2} < \cdots < \rho_{R,k_R}$ and $\rho_{T,1} < \rho_{T,2} < \cdots < \rho_{T,k_T}$. The even-numbered order statistics constrains the prior to lie within a certain range, where we constrain $\rho_R \in [-4,1]$ and $\rho_T \in [-1,2]$ (so the transition-transversion ratio is between 0.1 and 100). These intervals

where chosen to contain the log rates and log transition-transversion ratios that we expect to see in the alignment.

Due to the proposal moves we use with the Reversible Jump algorithm, we never have to calculate the value of even-numbered order prior directly. Instead, we only calculate the ratio of two prior settings, where one interval is added, removed or relocated. The ratio of the probabilities of a set of even-numbered order statistics intervals, to which an interval has been added is:

$$\frac{P([x_1,\ldots,x_i,x^*,x_{i+1},\ldots,x_k],k+1)}{P([x_1,\ldots,x_k],k)} = 2(2k+3)(k+1)\frac{(x_{i+1}-x^*)(x^*-x_i)}{x_{i+1}-x_i}, \tag{6.19}$$

where x^* is a new interval, which is situated between existing intervals x_i and x_{i+1} (Green, 1995). We define $E_{\min} = x_1$ and $E_{\max} = x_k$ as the upper and lower boundary of prior, which are not changed in our sampling scheme. Then, $x_2,\ldots x_{k-1}$ are used as the set of rates or transition-transversion ratios. Both the set of internal points $x_2,\ldots x_{k-1}$ and the number of points, k vary in our inference scheme. The ratio of the probability of the even-numbered order statistics when value x_i has been changed to x^*, where $x_{i-1} < x^* < x_{i+1}$ still holds is:

$$\frac{P([x_1,\ldots,x_{i-1},x^*,x_{i+1},\ldots,x_k],k+1)}{P([x_1,\ldots,x_k],k)} = \frac{(x_{i+1}-x^*)(x^*-x_{i-1})}{(x_{i+1}-x_i)(x_i-x_{i-1})}, \tag{6.20}$$

As a baseline to compare the other priors with, we also investigate a simple uniform distribution $\mathcal{U}[E_{\min},E_{\max}]$ over the range of the even-numbered order statistics, where:

$$\mathcal{U}[x;a,b] = \frac{\mathbb{I}(a \le x \le b)}{b-a} \tag{6.21}$$

is the uniform distribution over the interval $[a,b]$. We will often drop x from the specification of the distribution, as it does not effect the probability as long as it is within in the interval. Hence, our prior on the $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ both have the same form as Equation (6.16), with $Q_R(\rho_{R,i}) = \mathcal{U}[-4,1]$ and $Q_T(\rho_{T,i}) = \mathcal{U}[-1,2]$.

In summary, the full prior distribution for the model is:

$$\begin{aligned}\mathcal{P} = P(\boldsymbol{\nu},k_R,k_T,\boldsymbol{\rho}_R,\boldsymbol{\rho}_T|k_S,\boldsymbol{\rho}_S,\mathbf{C}) = P\left(\nu_S|k_S,C_S^{\min},C_S^{\max}\right)\times \\ \textstyle\prod_{A\in\{R,T\}} P(k_A)P(\boldsymbol{\rho}_A|k_A)P\left(\nu_A|k_A,C_A^{\min},C_A^{\max}\right),\end{aligned} \tag{6.22}$$

as defined in Equations (6.12), (6.15) and (6.16). We do not define prior distributions over k_S and $\boldsymbol{\rho}_S$ as these parameters are not changed within our model, owing to the fact that the number of different tree topologies is fixed.

6.4.3 Likelihood

Our *complete-data* likelihood is:

$$\begin{aligned}\mathcal{L} = P(\mathcal{D},\mathbf{h}|\boldsymbol{\nu},k_S,k_R,k_T,\boldsymbol{\rho}_S,\boldsymbol{\rho}_R,\boldsymbol{\rho}_T) = \textstyle\prod_{t=1}^{N} P(\mathbf{y}_t|H_{S,t},H_{R,t},H_{T,t})\times \\ \textstyle\prod_{A\in\{S,R,T\}}\left\{P(H_{A,1}|k_A,\boldsymbol{\rho}_A)\prod_{t=2}^{N}P(H_{A,t}|H_{A,t-1},\nu_A,k_A,\boldsymbol{\rho}_A)\right\},\end{aligned} \tag{6.23}$$

as defined in Equations (6.6), (6.9) and (6.10).

6.4.4 Posterior inference

In the Bayesian paradigm, we are interested in the posterior distribution of the parameters and the hidden variables:

$$P(\mathbf{h},\boldsymbol{\nu},k_R,k_T,\boldsymbol{\rho}_R,\boldsymbol{\rho}_T|\mathcal{D},k_S,\boldsymbol{\rho}_S,\mathbf{C}) \propto \mathcal{P}\times\mathcal{L} \tag{6.24}$$

where the prior $\mathcal{P}$ and likelihood $\mathcal{L}$ are defined in Equations (6.22) and (6.23). Recall that for certain bacteria and viruses the tree topology along the alignment can change as a consequence of recombination. This corresponds to a state transition $H_{S,t}=\rho_{S,i}\rightarrow H_{S,t+1}=\rho_{S,(k\neq i)}$ at the breakpoint t of the affected region. Likewise, different segments of a DNA sequence alignment can be under different selective pressure, which corresponds to transitions between different rate states $H_{R,t}$. Hence, our main objective is the prediction of the marginal posterior probabilities:

$$P(H_{A,t}|\mathcal{D},k_S,\boldsymbol{\rho}_S) \;=\; \sum_{H_{A,1}}\cdots\sum_{H_{A,t-1}}\sum_{H_{A,t+1}}\cdots\sum_{H_{A,N}} P(\mathbf{H}_A|\mathcal{D},k_S,\boldsymbol{\rho}_S) \tag{6.25}$$

where again $A\in\{S,R,T\}$, and the dependence on $\mathbf{C}$ has not been made explicit to simplify the notation. Plotting the distributions of $H_{S,t}$, $H_{R,t}$ and $H_{T,t}$ along the DNA sequence alignment gives clear indications about the location of recombinant regions, differently diverged regions and regions with changes in the transition-transversion ratio respectively. The distributions $P(\mathbf{H}_A|\mathcal{D},k_S,\boldsymbol{\rho}_S)$ are obtained by marginalisation of the posterior:

$$P(\mathbf{H}_A|\mathcal{D},k_S,\boldsymbol{\rho}_S)=\sum_{\mathbf{H}_{\{S,R,T\}\setminus A}}\sum_{k_R}\sum_{k_T}\int d\boldsymbol{\rho}_R\int d\boldsymbol{\rho}_T\int d\boldsymbol{\nu}P(\mathbf{h},\boldsymbol{\nu},k_R,k_T,\boldsymbol{\rho}_R,\boldsymbol{\rho}_T|\mathcal{D},k_S,\boldsymbol{\rho}_S) \tag{6.26}$$

While the marginalisation in Equation (6.25) can be carried out efficiently with linear time complexity using dynamic programming techniques discussed in Rabiner (1989), the implicit marginalisations to make Equation (6.24) into a distribution, and the explicit marginalisations in Equation (6.26) are intractable and have to be numerically approximated with Markov chain Monte Carlo (MCMC). For fixed $\boldsymbol{\rho}_R$ and k_R, Husmeier (2005) demonstrated a Gibbs sampling procedure (Casella and George, 1992) for $\mathbf{H}_S$, $\mathbf{H}_R$, ν_S and ν_R (the transition-transversion ratio was assumed to be invariant along the alignment). However, it is computationally intractable to directly sample from the appropriate marginal distributions of $\boldsymbol{\rho}_R$, k_R, $\boldsymbol{\rho}_T$ or k_T. The novelty of our approach comes from extending Husmeier (2005) to rigorously marginalise over $\boldsymbol{\rho}_R$, $\boldsymbol{\rho}_T$, k_R and k_T by adopting a Reversible Jump Metropolis-Hastings scheme (Green, 1995), allowing us to generate samples for $\boldsymbol{\rho}_R$, k_R, $\boldsymbol{\rho}_T$ and k_T despite the dimensionality of the parameter space changing. The motivation for our model comes from Boys and Henderson (2004), who applied RJMCMC to inference in non-phylogenetic HMMs for segmenting individual DNA sequences. We refer to our model, which generalises this approach to the segmentation of whole DNA sequence alignments in a phylogenetic context, as the Phylogenetic Reversible Jump Factorial Hidden Markov model (PRJ-FHMM).

6.4.5 The posterior probability that ν_R and ν_T are not relevant

As the number of rate states can vary, the model may find that the alignment is best modelled by a single rate state instead of multiple rate states. This implies that ν_R has no effect on the system, and is hence irrelevant. We calculate the posterior probability of ν_R being relevant as:

$$M_R = \frac{1}{N}\sum_i \mathbb{I}\left(k^{(i)} > 1\right). \tag{6.27}$$

where $k_R^{(i)}$ is the i^{th} sample of k_R. We then say that the posterior probability of ν_R being relevant is M_R. Correspondingly, the probability of ν_R not being relevant, and the alignment being better modelled by a single rate state is $1 - M_R$. The same argument applies to ν_T and log transition-transversion ratio states. ν_S is always relevant or irrelevant depending on how the topology states $\boldsymbol{\rho}_S$ are fixed.

For all plots of the posterior distribution of $P(\nu_R|\mathcal{D})$ and $P(\nu_T|\mathcal{D})$, we will mention the posterior probability of ν_R and ν_T being relevant to modelling the alignment.

6.5 Inference using Reversible Jump MCMC

The following moves are performed for each iteration of the MCMC sampler:

1. Sample $\mathbf{H}_A \sim P\left(\cdot|\nu_A, \boldsymbol{\rho}_A, k_A, \mathbf{H}_{\{S,R,T\}\setminus A}, \mathcal{D}\right)$ for $A = \{S,R,T\}$.
2. Sample $\nu_A \sim P\left(\cdot|C_A^{\min}, C_A^{\max}, k_A, \mathbf{H}_A, \mathcal{D}\right)$ for $A = \{S,R,T\}$.
3. For $A = \{R,T\}$:
   (a) Propose $\boldsymbol{\rho}_A^*$ and k_A^* by adapting $\boldsymbol{\rho}_A$ and k_A.
   (b) Propose $\mathbf{H}_A^* \sim P\left(\cdot|\nu_A, \boldsymbol{\rho}_A^*, k_A^*, \mathbf{H}_{\{S,R,T\}\setminus A}, \mathcal{D}\right)$.
   (c) Accept $\boldsymbol{\rho}_A^*$, k_A^* and $\mathbf{H}_A^*$ if $\mathcal{U}[0,1] <$ acceptance probability, where $\mathcal{U}[0,1]$ is a sample from the uniform distribution over the unit interval.

Note that the conditioning part of each distribution contains the Markov blanket (Pearl, 1988) of the respective random variable to be sampled. Recall that a random variable given its Markov blanket is independent from the remaining variables in the domain. Hence, conditioning on the Markov blanket is equivalent to conditioning on the complete set of random variables (without the sampled one). The Markov blanket of each random variable can easily be read off from Figure 6.1b: B is in A's Markov blanket if and only if there is either an edge between A and B, or both A and B are parents of another random variable (Pearl, 1988). This proves that the proposed scheme is a valid Gibbs sampling scheme.

Note that, in principle, π_A, π_C, π_G and π_T (the equilibrium nucleotides frequencies) from θ in Equation (6.6) should be included in the Gibbs sampling scheme, as described in Husmeier

and McGuire (2003). However, Husmeier and McGuire (2003) found that in practice a fixation of π_A, π_C, π_G and π_T at values estimated from their occurrences in the alignment makes little difference and has the advantage of reduced computational costs.

6.5.1 Sampling $\mathbf{H}_A \sim P\left(\cdot|\nu_A, \boldsymbol{\rho}_A, k_A, \mathbf{H}_{\{S,R,T\}\setminus A}, \mathcal{D}\right)$

Sampling the hidden state sequences $\mathbf{H}_S$, $\mathbf{H}_R$ and $\mathbf{H}_T$ can be effected with a Gibbs-within-Gibbs procedure, as described in Husmeier and McGuire (2003). However, the stochastic forward-backward algorithm of Boys et al. (2000) has proven to lead to a faster mixing and convergence of the Markov chain (Werhli et al., 2006) and was, thus, used in the simulations reported in this chapter.

6.5.2 Sampling $\nu_A \sim P\left(\cdot|C_A^{\min}, C_A^{\max}, k_A, \boldsymbol{\rho}_A, \mathbf{H}_A, \mathcal{D}\right)$

The sampling steps for ν_S, ν_R and ν_T are straightforward due to the conjugacy of the beta distribution $\mathcal{B}$, as defined in Equation (6.11). Define:

$$\Psi_A = \sum_{t=1}^{N-1} \mathbb{I}\left(H_{A,t} = H_{A,t+1}\right) \qquad \overline{\Psi}_A = N - 1 - \Psi_A \tag{6.28}$$

It is then easy to show from (6.9) that:

$$P\left(\nu_A|C_A^{\min}, C_A^{\max}, \mathbf{H}_A, k_A, \mathcal{D}\right) \propto \mathbb{I}\left(C_A^{\min} \leq \nu_A \leq C_A^{\max}\right) \mathcal{B}\left(\nu_A|\Psi_A + \alpha, \overline{\Psi}_A + \beta\right). \tag{6.29}$$

See Husmeier and McGuire (2003) for a derivation for the untruncated case. If $k_A = 1$, then Equation (6.29) does not apply, as there is only a single possible $\mathbf{H}_A$. Given that sampling from the beta distribution is straightforward (see, e.g. Rubinstein, 1981), we can easily generate samples from the truncated beta distribution by repeatedly sampling ν from the beta distribution until we get a sample satisfying $\nu > C$. In practice however, the resulting acceptance probability can be too low. Instead, we sample these parameters by performing 50 Metropolis-Hastings samples, proposed uniformly between $C^{\min}$ and $C^{\max}$, followed by 200 Metropolis-Hastings steps proposed from a Gaussian centred on the current ν value, with the standard deviation set to the distance to the closest boundary. While even a single Metropolis-Hastings step would still allow the model to converge eventually, in practice sampling from this one dimensional space is computationally cheap. The final sample produced from this scheme was found to be sufficiently uncorrelated for minimising the overall computational cost of our inference scheme.

Additionally, Ψ_A and $\overline{\Psi}_A$ allow us to generate an estimate of the posterior distribution of ν_A:

$$P\left(\nu_A|\mathcal{D}, C_A^{\min}, C_A^{\max}\right) \approx \frac{1}{S}\sum_i \frac{\mathcal{B}\left(\nu_A|\Psi_A^{(i)} + \alpha, \overline{\Psi}_A^{(i)} + \beta\right)}{\int_{C_A^{\min}}^{C_A^{\max}} \mathcal{B}\left(\nu_A|\Psi_A^{(i)} + \alpha, \overline{\Psi}_A^{(i)} + \beta\right) d\nu_A} \mathbb{I}\left(C_A^{\min} \leq \nu_A \leq C_A^{\max}\right) \tag{6.30}$$

where the superscript (i) represents the i^{th} sample and we have S samples. The integral is easily calculated using the trapezoid method.

6.5.3 Proposing and conditionally accepting $\boldsymbol{\rho}_A^*$, k_A^* and $\mathbf{H}_A^*$

We adopt a Reversible Jump Metropolis-Hastings scheme (Green, 1995) where we propose a new number of rate states k_R^* and a new set of rate states $\boldsymbol{\rho}_R^*$ from k_R and $\boldsymbol{\rho}_R$. This is done using a birth move (with probability b_k), a death move (with probability d_k) or a relocation of one of the rate states (with probability r_k). A new $\mathbf{H}_R^*$ is then proposed given the new $\boldsymbol{\rho}_R{}^*$. The new set of rate states $\boldsymbol{\rho}_R^*$ is then accepted with a probability such that given ergodicity, the Markov chain is guaranteed to converge in distribution to the correct posterior distribution. This procedure is similar to the reversible jump move (b) from Boys and Henderson (2004).

$\boldsymbol{\rho}_T$ and k_T are adapted in the same way with identical derivations, so we only show the derivation for $\boldsymbol{\rho}_R$ and drop the R subscript on k_R. To use the Reversible Jump method, we need to specify how we propose k^* and $\boldsymbol{\rho}_R^*$.

The set of all possible proposal moves is outlined in Table 6.1 for the Gaussian or uniform prior distribution, and Table 6.2 for the even-numbered order statistics prior. If using an ordered prior like the even-numbered order statistics prior, the rate states in $\boldsymbol{\rho}_R^*$ are then sorted to satisfy the ordering constraints. Note that k^* is proposed such that the Hastings factor cancels out against the prior ratio. Lastly, we propose $\mathbf{H}_R^* \sim P(\cdot|\nu_R, \boldsymbol{\rho}_R^*, k^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D})$ as described in Section 6.5.1.

The acceptance probability a of k^*, $\boldsymbol{\rho}_R^*$ and $\mathbf{H}_R^*$ is $\min\{1, A_B\}$, where:

$$A_B = \text{Likelihood ratio} \times \text{Prior ratio} \times \text{Inverse proposal probability ratio} \times |\det(\text{Jacobian})|, \tag{6.31}$$

see Green (1995) – our formulation is closer to that of Suchard et al. (2003). We first derive the acceptance probability of a birth move. We first propose a new rate state $\rho_R{}^*$ from Q_R in Equation (6.16). We then map $(\boldsymbol{\rho}_R, \rho_R^*)$ to $(\boldsymbol{\rho}_R^*)$. If using an ordered prior such as the even-numbered order statistic prior, this map is such that $\boldsymbol{\rho}_R^*$ is sorted. In Equation (6.31), the Jacobian term refers to this mapping, and det stands for the determinant. This mapping is a permutation, hence the Jacobian is a permutation matrix, which implies $\det(\text{Jacobian}) = \pm 1$, so $|\det(\text{Jacobian})| = 1$.

Move type	Probability of move and proposal for $\boldsymbol{\rho}_R{}^*$	Description of how $\boldsymbol{\rho}_R{}^*$ is proposed
Birth $k^* = k+1$	$b_k = c\min\left\{1, \frac{P(k+1)}{P(k)}\right\}$ $\pi_b(\boldsymbol{\rho}_R{}^*\|\boldsymbol{\rho}_R) = \frac{1}{k+1}Q(\rho_R{}^*)$	A new rate is sampled from Q in Equation (6.16), the prior distribution on $\boldsymbol{\rho}_R$ for a single rate. Where to insert the new rate state is randomly and uniformly sampled from the $k+1$ possibilities.
Death $k^* = k-1$	$d_k = c\min\left\{1, \frac{P(k-1)}{P(k)}\right\}$ $\pi_d(\boldsymbol{\rho}_R{}^*\|\boldsymbol{\rho}_R) = \frac{1}{k}$	A randomly chosen rate is deleted.
Relocation $k^* = k$	$r_k = 1-(b_k+d_k)$ $\pi_r(\boldsymbol{\rho}_R{}^*\|\boldsymbol{\rho}_R) = \frac{1}{k}Q(\rho_R{}^*)$	An existing rate factor position is randomly chosen, and its position resampled from Q (see birth move).

Table 6.1: Possible proposal moves for a uniform or Gaussian prior, the probability with which they are selected, and the corresponding proposal probability $\pi_M(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R)$ for $\boldsymbol{\rho}_R{}^*$. All π distributions presume that $\boldsymbol{\rho}_R{}^*$ is a valid proposal given the move type, as otherwise the π distributions are not normalised. We use $c = 0.4$ – see Green (1995).

6.5.4 Acceptance probability with a uniform or Gaussian prior on $\boldsymbol{\rho}_R$

From Equations (6.16), (6.22) and (6.23), and Table 6.1 we see that after cancelling, the terms (by their initials) are:

$$\begin{aligned}
\text{LR} &= \frac{P(\mathcal{D}|\mathbf{H}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|\mathbf{H}_R,\mathbf{H}_S,\mathbf{H}_T)}\frac{P(\mathbf{H}_R^*|\mathcal{D},k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathbf{H}_R|\mathcal{D},k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \\
&= \frac{P(\mathcal{D},\mathbf{H}_R{}^*|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D},\mathbf{H}_R|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \\
&= \frac{P(\mathbf{H}_R^*|\mathcal{D},k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathbf{H}_R|\mathcal{D},k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)}\frac{P(\mathcal{D}|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \qquad (6.32) \\
\text{PR} &= \frac{P(k+1)}{P(k)}\frac{P(\boldsymbol{\rho}_R^*|k+1)}{P(\boldsymbol{\rho}_R|k)} = \frac{P(k+1)}{P(k)}[Q(\rho_R^*)] \qquad (6.33) \\
\text{IPPR} &= \frac{P(\mathbf{H}_R|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}{P(\mathbf{H}_R{}^*|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}\frac{d_{k+1}\pi_d(\boldsymbol{\rho}_R|\boldsymbol{\rho}_R{}^*)}{b_k\pi_b(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R)} \\
&= \frac{P(\mathbf{H}_R|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}{P(\mathbf{H}_R{}^*|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}\frac{P(k)}{P(k+1)}\frac{k+1}{k+1}\frac{1}{Q(\rho_R{}^*)} \qquad (6.34)
\end{aligned}$$

PR cancels against IPPR, except for the proposal probability ratio for $\mathbf{H}_R$ which in turn cancels with ratio of $P(\mathbf{H}_R|\mathcal{D},\ldots)$ in LR. Hence:

$$A_B = \frac{P(\mathcal{D}|k+1,\nu_R,\boldsymbol{\rho}_R{}^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)}, \qquad (6.35)$$

where due to the HMM structure, $P(\mathcal{D}|\boldsymbol{\rho}_R{}^*, k+1, \mathbf{H}_S, \mathbf{H}_T, \nu_R)$ can be computed from the complete likelihood of Equation (6.23) in linear time with a dynamical programming algorithm known as the forward algorithm (Rabiner, 1989). Note that the stated dependence on the conditioning variables becomes clear from the conditional independence graph of Figure 6.1b and the properties of the Markov blanket, as discussed above. Using HMMs has had two main benefits for our model: efficient sampling from the marginal distribution of $\mathbf{H}_R$, and efficient integrating over all possible $\mathbf{H}_R$.

The same cancellations and simplifications occur when considering the acceptance probability of the death move as the death move is the inverse of the birth move. Hence the acceptance probability of a death move is:

$$A_D = \frac{P(\mathcal{D}|k, \nu_R, \boldsymbol{\rho}_R{}^*, \mathbf{H}_S, \mathbf{H}_T)}{P(\mathcal{D}|k+1, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T)}, \tag{6.36}$$

which is identical to Equation (6.35) after replacing $k^* = k+1$ with $k^* = k-1$. The acceptance probability of a relocation is:

$$A_R = \frac{P(\mathcal{D}|k, \nu_R, \boldsymbol{\rho}_R{}^*, \mathbf{H}_S, \mathbf{H}_T)}{P(\mathcal{D}|k+1, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T)}, \tag{6.37}$$

which is again the same to also the same Equation (6.35) after replacing $k^* = k+1$ with $k^* = k$ as relocation moves are symmetrical to themselves, in the same way as the birth and death moves are symmetrical.

Note that Equation (6.29) does not apply when $k = 1$ as there is only a single possible $\mathbf{H}_R$. Hence ν_R has no effect on the likelihood. When moving from two rates to a single rate state, ν_R is removed from the system. Correspondingly, when moving from a single rate state to two rate states, ν_R is proposed from the prior. To see that this leaves the acceptance ratios unchanged, first consider the death move from two rate states to a single rate. We have an extra $P\left(\nu_R|C_R^{\min}, C_R^{\max}\right)$ in the denominator of PR, and a new proposal term $Q(\nu_R)$ in the numerator of the IPPR. We set $Q(\nu_R) = P\left(\nu_R|C_R^{\min}, C_R^{\max}\right)$ so that these terms cancel, leaving the acceptance probability unchanged. The reverse argument applies to the birth move, so the acceptance probability is again unchanged.

6.5.4.1 Acceptance probability with the even-numbered order statistics on $\boldsymbol{\rho}_R$

When using even-numbered order statistics, the acceptance probability is different as the prior and inverse proposal probability ratios for $\boldsymbol{\rho}_R$ are no longer symmetrical. This is because ρ_R^* is proposed from R, a uniform distribution that covers the interval of the even-ordered number statistics and not directly from the prior.

Referring to Table 6.2 and Equation (6.19), we see that the prior ratio PR_e for the even-

Move type	Probability of move and proposal for $\boldsymbol{\rho}_R^*$	Description of how $\boldsymbol{\rho}_R^*$ is proposed
Birth $k^* = k+1$	$b_k = c\min\left\{1, \frac{P(k+1)}{P(k)}\right\}$ $\pi_b(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R) = \frac{1}{E_{\max}-E_{\min}}$	A new rate is sampled from $\mathcal{U}[E_{\min}, E_{\max}]$, a uniform distribution over the range of the even-numbered order statistics prior.
Death $k^* = k-1$	$d_k = c\min\left\{1, \frac{P(k-1)}{P(k)}\right\}$ $\pi_d(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R) = \frac{1}{k}$	A randomly chosen rate is deleted.
Relocation $k^* = k$	$r_k = 1-(b_k+d_k)$ $\pi_r(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R) = \frac{1}{k}\frac{1}{E_{\max}-E_{\min}}$	An existing rate factor position is randomly chosen, and its position re-sampled from $\mathcal{U}[E_{\min}, E_{\max}]$ (see birth move).

Table 6.2: Possible proposal moves for the even-numbered order statistics prior, the probability with which they are selected, and the corresponding proposal probability $\pi_M(\boldsymbol{\rho}_R^*|\boldsymbol{\rho}_R)$ for $\boldsymbol{\rho}_R{}^*$. All π distributions presume that $\boldsymbol{\rho}_R^*$ is a valid proposal given the move type, as otherwise the π distributions are not normalised. We use $c = 0.4$ – see Green (1995).

numbered order statistics prior is:

$$\begin{aligned}\mathrm{PR}_e &= \frac{P(\mathbf{H}_R^*|k+1,\nu_R)}{P(\mathbf{H}_R|k,\nu_R)}\frac{P(\boldsymbol{\rho}_R^*|k+1)}{P(\boldsymbol{\rho}_R|k)}\frac{P(k+1)}{P(k)} \\ &= \frac{P(\mathbf{H}_R^*|k+1,\nu_R)}{P(\mathbf{H}_R|k,\nu_R)}2(2k+3)(k+1)\frac{(\rho_{i+1}-\rho^*)(\rho^*-\rho_i)}{\rho_{i+1}-\rho_i}\frac{P(k+1)}{P(k)} \qquad (6.38)\end{aligned}$$

where ρ^* lies between existing rates $\rho_{i-1} < \rho^* < \rho_i$. ρ_{i-1} can be the lower bound, and ρ_{i+1} can be the upper bound on the even-numbered statistics prior. Again, referring back to Table 6.2, we see that the inverse proposal ratio IPPR_e for the even-numbered order statistics prior is:

$$\begin{aligned}\mathrm{IPPR}_e &= \frac{P(\mathbf{H}_R|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}{P(\mathbf{H}_R{}^*|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}\frac{d_{k+1}\pi_d(\boldsymbol{\rho}_R|\boldsymbol{\rho}_R^*)}{b_k\pi_b(\boldsymbol{\rho}_R^*|\boldsymbol{\rho}_R)} \\ &= \frac{P(\mathbf{H}_R|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}{P(\mathbf{H}_R{}^*|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}\frac{P(k)}{P(k+1)}\frac{1}{k+1}\frac{1}{E_{\max}-E_{\min}}. \qquad (6.39)\end{aligned}$$

Multiplying Equations (6.32), (6.38) and (6.39) gives:

$$\begin{aligned}\mathrm{LR}\times\mathrm{PR}_e\times\mathrm{IPPR}_e &= \frac{P(\mathbf{H}_R^*|\mathcal{D},k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathbf{H}_R|\mathcal{D},k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)}\frac{P(\mathcal{D}|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \\ &\quad\times\frac{P(\mathbf{H}_R|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})}{P(\mathbf{H}_R^*|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T,\mathcal{D})} \\ &\quad\times 2\frac{(2k+3)}{E_{\max}-E_{\min}}\frac{(\rho_{i+1}-\rho^*)(\rho^*-\rho_i)}{\rho_{i+1}-\rho_i}\frac{P(k+1)}{P(k)}\frac{P(k)}{P(k+1)}\frac{k+1}{k+1}\end{aligned}$$

where the terms involving $P(k)$, $P(k+1)$, $k+1$ or $\mathbf{H}_R$ cancel. Simplifying and substituting

back into Equation (6.31) gives an overall acceptance probability of:

$$A_B^e = \frac{P(\mathcal{D}|k+1,\nu_R,\boldsymbol{\rho}_R{}^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \frac{2(2k+3)}{E_{\max}-E_{\min}} \frac{(\rho_{i+1}-\rho^*)(\rho^*-\rho_i)}{\rho_{i+1}-\rho_i}. \tag{6.40}$$

Again, the inverse proposal move of the birth move is a death move, so the derivation is almost unchanged. Terms involving $\mathcal{D}$ and $\mathbf{H}_R$ simplify as shown in the derivation for the birth move, while swapping terms that model k^* and $\boldsymbol{\rho}_R^*$ between the prior and the inverse proposal gives an acceptance probability for a death move of:

$$A_D^e = \frac{P(\mathcal{D}|k+1,\nu_R,\boldsymbol{\rho}_R{}^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \frac{E_{\max}-E_{\min}}{2(2k+3)} \frac{\rho_{i+1}-\rho_i}{(\rho_{i+1}-\rho^*)(\rho^*-\rho_i)}. \tag{6.41}$$

For the derivation of the relocation move, terms involving $\mathcal{D}$ and $\mathbf{H}_R$ simplify as shown in the derivation for the birth move. Inserting Equation (6.20) into the prior ratio, we see that the acceptance probability is:

$$A_R^e = \frac{P(\mathcal{D}|k+1,\nu_R,\boldsymbol{\rho}_R^*,\mathbf{H}_S,\mathbf{H}_T)}{P(\mathcal{D}|k,\nu_R,\boldsymbol{\rho}_R,\mathbf{H}_S,\mathbf{H}_T)} \frac{(\rho_{i+1}-\rho^*)(\rho^*-\rho_{i-1})}{(\rho_{i+1}-\rho^i)(\rho^i-\rho_{i-1})}, \tag{6.42}$$

where ρ^* is the new rate sampled for the old rate state ρ_i - see Table 6.2. See Section 6.5.4 for a discussion of why the acceptance probability are not changed when $k=1$ or $k^*=1$.

6.5.5 Checking the correctness of the implementation of the inference scheme

Three separate functions of the code were tested for correctness: calculating the probability $P(\mathbf{y}_t|H_{S,t},H_{R,t},H_{T,t},\pi)$, generating sequence alignments and the reversible jump inference scheme outlined in Section 6.5.

To test our calculation of $P(\mathbf{y}_t|H_{S,t},H_{R,t},H_{T,t},\pi)$, we generated a variety of synthetic sequence alignments of known topologies, rate, and transition-transversion ratio using the program SEQ-GEN (Rambaut and Grassly, 1997). Following the assumption in Equation (6.1) – which we know holds in this case – we calculated likelihood values for a wide range of rates and transition-transversion ratios. This technique identified mismatches between the definitions of the rate and transition-transversion ratio that were used by DNAML (Felsenstein, 1981) and SEQ-GEN (Rambaut and Grassly, 1997) versus those initially used by our model.

In order to check the correctness of our code for generating synthetic DNA sequence alignments, we generated a variety of sequences alignments from our implementation, and checked that the rate and transition-transversion ratio of the resulting alignments were correctly estimated by DNAML (Felsenstein, 1981). Hence, we can be reasonably certain that both our method to generate sequence alignments, and our method for estimating the probability of a given column in the alignment are correct.

Finally, we checked the correctness of the reversible jump MCMC scheme. The topology and transition-transversion ratio were set to fixed values, which sufficiently reduces the complexity of the problem that numerical integration can be used to calculate interesting posterior

distributions directly. In particular, we used a simple trapezium rule to approximate the posterior distributions for one or two rate states being present in the alignment. This is only feasible due to being able to efficiently marginalise over $\mathbf{H}_R$ with the forwards backwards algorithm to get the likelihood of the alignment.

The likelihood of the data was evaluated at set of ρ and ν_R values. The set of ρ values was $\mathcal{R} = (-4, -3.95, -3.9, \ldots, 0.5)$ while the set of ν_R was: $\mathcal{N} = (0, 0.005, 0.01, \ldots, 1)$. Then:

$$P(\boldsymbol{\rho}_R = \{r\} \,|\mathcal{D}, k_R = 1) \approx \frac{1}{Z_1} P(\mathcal{D}|k_R = 1, \boldsymbol{\rho}_R = \{r\}) P(\boldsymbol{\rho}_R = \{r\} \,|k_R = 1), \tag{6.43}$$

where

$$Z_1 = \frac{(\max(\mathcal{R}) - \min(\mathcal{R}))}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} P(\mathcal{D}|k_R = 1, \boldsymbol{\rho}_R = \{r\}) P(\boldsymbol{\rho}_R = \{r\} \,|k_R = 1).$$

Z_1 is the normalising constraint, which takes into account the density of the points at which we evaluate our approximation to the posterior. This is actually an approximation to the trapezoid method, and assumes that points in $\mathcal{R}$ are evenly spaced. The posterior distribution of the variables of interest given there are two rate states is:

$$P(\boldsymbol{\rho}_R = \{l, h\}, \nu_R = n | \mathcal{D}, k_R = 2) \approx \frac{1}{Z_2} P(\mathcal{D}|k_R = 1, \boldsymbol{\rho}_R = \{l, h\}, \nu_R = n) P(\boldsymbol{\rho}_R = \{l, h\} \,|k_R = 2) P(\nu_R = n), \tag{6.44}$$

where:

$$Z_2 = \frac{1/2(\max(\mathcal{R}) - \min(\mathcal{R}))^2}{1/2(|\mathcal{R}| + 2)(|\mathcal{R}| + 1)} \frac{(\max(\mathcal{N}) - \min(\mathcal{N}))}{|\mathcal{N}|} \times \sum_{n \in \mathcal{N}} \sum_{l \in \mathcal{R}} \sum_{h \in \mathcal{R}} P(\mathcal{D}|k_R = 1, \boldsymbol{\rho}_R = \{l, h\}, \nu_R = n) P(\boldsymbol{\rho}_R = \{l, h\} \,|k_R = 2) P(\nu_R = n).$$

From Equation (6.44), we can easily calculate the individual posterior distributions for ν_R and $\boldsymbol{\rho}_R$ by marginalisation. Additionally, from the normalisation constants, we can compute the Bayes factor for two rate states as opposed to one rate state begin present in the alignment:

$$\frac{P(k_R = 2|\mathcal{D})}{P(k_R = 1|\mathcal{D})} \approx \frac{P(k_R = 2)}{P(k_R = 1)} \times \frac{Z_2}{Z_1}.$$

We compared the results for this numerical interaction scheme and our reversible jump inference scheme on the Neisseria alignment described in Section 6.8.1, All posterior distributions were found to closely match between those calculated with this method and those inferred using our RJMCMC scheme. Overall, this should give a high confidence that the model is correctly implemented.

6.6 Setting up the simulations

6.6.1 Specific Markov chain settings and convergence diagnostics

We used the method of Gelman and Rubin (1992) to check for convergence by computing the Potential Scale Reduction Factors (PSRF) of $H_{R,t}$ and $H_{T,t}$ for $t \in \{1,\ldots,N\}$, and ν_A. These characteristics were chosen as they are invariant to the dimensionality of the parameter space. All results presented in this chapter, for all models, were run in triplicate, with the exception of the synthetic codon effect study. In the synthetic codon effect study, we started ten simulations with a uniform spread of initial ν_R values.

For our proposed model, the initial number of rates was picked uniformly between 1 and $k_{\max}$, with each rate sampled randomly from the uniform distribution. In this thesis, we are mainly interested in investigating the rate along the alignment, so all runs were started with only a single transition-transversion ratio, randomly sampled from the uniform distribution.

We discarded the first 10,000 iterations of the PRJ-FHMM samples as the burn-in period. Then, for the next 200,000 iterations every 10^{th} sample was kept so that we could form the posterior summaries. The Multiple-Changepoint Process (MCP) of Minin et al. (2005) was run for 200,000 burn-in iterations, followed by 4,000,000 sampling iterations where every 200th sample was kept. These lengths were chosen as they resulted in sufficient convergence indications, as measured by the PSRF. The highest PSRF observed for any characteristic for the PRJ-FHMM was 1.0015, while for the MCP it was 1.077, indicating a sufficient degree of convergence. The maximum PSRF for the MCP resulted from the runs with a low value of λ_R^B, but these values are still sufficiently small to indicate sufficient convergence.

6.6.2 Comparisons with a breakpoint model

Suchard et al. (2003) introduced the Multiple-Change Point (MCP) model where a DNA sequence alignment is split into discrete segments by a series of breakpoints. The number of segments is thus always one greater than the number of breakpoints. The topology, rate and transition-transversion ratio are estimated independently between each successive pair of breakpoints. This joint estimation causes *a priori* correlation between the locations of the changes in the rate and the locations of changes in the topology, which Minin et al. (2005) removed by using one MCP to model the topology changes, and a separate MCP to model changes in both the rate and transition-transversion ratio. A software implementation of their model is available from `http://www.biomath.ucla.edu/msuchard/DualBrothers/`.

Minin et al. (2005) place truncated Poisson distributions over b_R, the number of breakpoints of the rate or transition-transversion ratio along the alignment, and b_S, the number of

breakpoints of the topology along the alignment:

$$P\left(b_R|\lambda_R^B\right) \propto \frac{\left(\lambda_R^B\right)^{b_R}}{b_R!}\mathbb{I}(b_R < N) \qquad P\left(b_S|\lambda_S^B\right) \propto \frac{\left(\lambda_S^B\right)^{b_S}}{b_S!}\mathbb{I}(b_S < N) \tag{6.45}$$

where λ_R^B and λ_S^B define the *a priori* expected mean numbers of joint rate and transition-transversion ratio, and topology breakpoints respectively.

While the MCP of Minin et al. (2005) separates out estimating the phylogenetic topology, their model still jointly estimates the rate and transition-transversion ratio. The PRJ-FHMM estimates these quantities independently, complicating our theoretical comparison. For the purposes of comparing the models, we will henceforth assume that $\boldsymbol{\rho}_R$ in the PRJ-FHMM has been augmented to additionally contain the transition-transversion ratio associated with each rate, allowing us to ignore ν_T. This simplification of the PRJ-FHMM allows us to make the following observations. In the PRJ-FHMM, ν_R defines a binomial distribution over the number of rate breakpoints in the alignment:

$$P(b_R|\nu_R) = \binom{N-1}{b_R}(1-\nu_R)^{b_R}\nu_R^{(N-1)-b_R}. \tag{6.46}$$

where this argument also applies to ν_S (ν_T is assumed to have been merged into ν_R). In the limit of $N \to \infty$ and $\nu_R \to 1$, the distribution in Equation (6.46) tends to that of (6.45), with:

$$\nu_R = 1 - \frac{\lambda_R^B}{N-1}, \tag{6.47}$$

as the Poisson distribution is the limiting case of the binomial distribution (Poisson, 1837). In practice, these distributions are extremely similar for small values of λ_R^B. Hence, the Poisson priors used in the MCPs of Suchard et al. (2003) and Minin et al. (2005) can be regarded as almost equivalent to a single setting of ν_S and ν_R, the transition probabilities in the PRJ-FHMM. To summarise: the MCP of Minin et al. (2005) is effectively a special case of our model with a fixed value of ν_S and ν_R, each state occurring only once and no separation between the processes leading to changed rates and changed nucleotide substitution parameters. In contrast, our PRJ-FHMM separates k_R, the number of states (or different segment types) from the average segment length (determined by ν_R), where the average segment length is not set to a fixed value, but also inferred.

6.6.3 Generating synthetic alignments

Given an alignment position that has rate r, we sample the branch lengths from the gamma distribution: $\mathcal{G}\left(\cdot|\mu = 10^r, \sigma = 10^{-5}\right)$, where

$$\mathcal{G}(x|\mu,\sigma) = x^{\frac{\mu^2}{\sigma}-1}\left(\frac{\mu}{\sigma}\right)^{\frac{\mu^2}{\sigma}}\frac{\exp\left(-\frac{\mu}{\sigma}x\right)}{\Gamma\left(\frac{\mu^2}{\sigma}\right)} \tag{6.48}$$

is the Gamma distribution parametrised in terms of the mean μ and variance σ. Both the PRJ-FHMM model and the MCP of Minin et al. (2005) expect some variance in the branch lengths, as they integrate out the exact branch lengths using the exponential distribution in Equation (6.6). Instead of matching the variance introduced by the exponential distribution, we instead set the variance on the branch lengths to be 10^{-5}. This introduces a small model mismatch as we do not in general expect alignments to exactly match our model assumptions.

As the focus in our simulations is on modelling changes in $\boldsymbol{\rho}_R$, the rate, the mutations in the synthetic datasets were always sampled with the transition-transversion ratio set to 2.27, the estimated transition-transversion ratio for the Neisseria alignment. Our equilibrium frequencies were set to the uniform distribution, with the exception of the alignments generated for checking the correctness of our model in Section 6.5.5.

Our distribution over a column in the alignment is $P(\mathbf{y}_t|H_S,\mathbf{w},\theta)$. This depends on the topology, the branch lengths, the transition-transversion ratio and the equilibrium distribution of the nucleotides. We have specified all of these, and can now sample $\mathbf{y}_t$ from this distribution.

6.7 Investigating the behaviour of PRJ-FHMM

6.7.1 The advantages of adapting $\boldsymbol{\rho}_R$

In Figure 6.3, we demonstrate the need for adapting $\boldsymbol{\rho}_R$, the set of rate states. We generate a synthetic alignment where the rate states are picked to lie between the fixed rates used by Husmeier (2005), representing a worst-case scenario for the fixed-parameter model. The fixed set of rates are set to the default set of rate states in Husmeier (2005):

$$\boldsymbol{\rho}_R = \{-3,-2.5,-2,-1.5,-1,-0.5,0,0.5,1,2\}.$$

The rates used in the synthetic alignment were $\{-1.75,-1.25,-0.75\}$, which were used to generate the alignment as outlined in Section 6.6.3. We also include a segment where recombination occurs, modelled by a change in the topology. As the focus was on finding the effect of adapting $\boldsymbol{\rho}_R$, we fixed $\boldsymbol{\rho}_T = \{0.35\}$. This sets the transition-transversion ratio to the correct value for the alignment.

Both models map each position in the alignment to a rate in $\boldsymbol{\rho}_R$. Ideally, the fixed parameter model should find that the two rates surrounding the true rate are equally likely, as then the mean predicted rate would equal the true underlying rate. Instead, as the model moves along the alignment, it consistently and frequently changes between the two closest rate states, poorly modelling the rate and causing spurious predictions of rate variation. When $\boldsymbol{\rho}_R$ is adapted, the model successfully finds the underlying rate along the alignment, clearly demonstrating the need for adapting $\boldsymbol{\rho}_R$. The fixed-parameter model could still correctly distinguish between the recombination and rate variation.

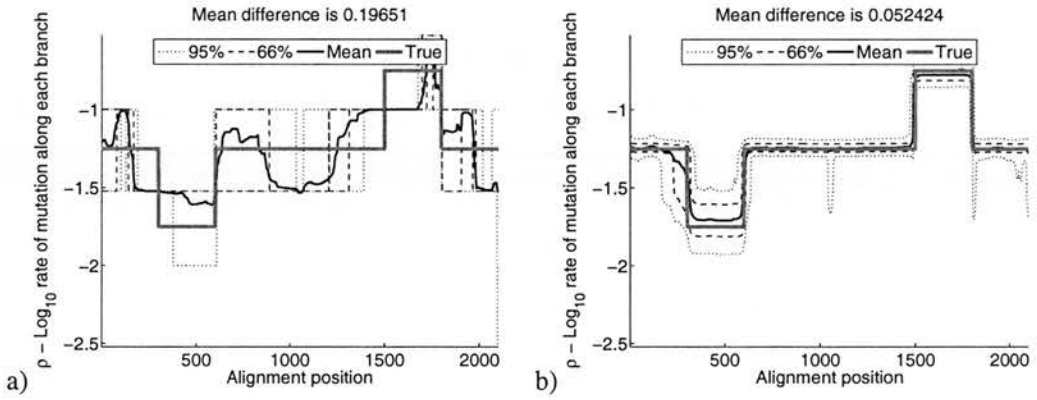


Figure 6.3: A synthetic alignment that causes problems for the fixed-parameter phylogenetic FHMM. Each segment has a rate picked from: $\{-1.75, -1.25, -0.75\}$. Sub-figure a) shows the credibility intervals of the posterior distribution of the rate for the fixed parameter model of Husmeier (2005) with the default set of possible rates: $\boldsymbol{\rho}_R = \{-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 2\}$. Sub-figure b) shows the credibility intervals of the posterior distribution of the rate when $\boldsymbol{\rho}_R$ is adapted. In both graphs, the rate is represented by the vertical axis, while the position along the horizontal axis represents the position along the alignment. The solid thick grey line shows the true underlying rate, while the solid black line shows the mean predicted rate and the dashed and dotted lines show the 66% and 95% percentiles of the posterior rate distribution.

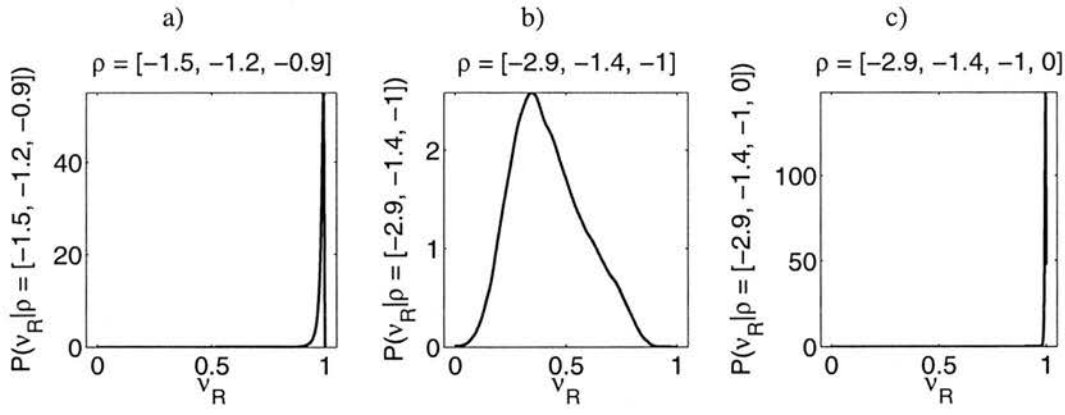


Figure 6.4: The influence of $\boldsymbol{\rho}_R$ on the posterior of ν_R, demonstrated on the alignment of Neisseria which will be described in Section 6.8.1. Sub-figures a) and b) illustrate that the posterior of ν_R is dependent on the choice of $\boldsymbol{\rho}_R$. The expected value of ν_R in Sub-figure a) is close to 1, which implies that the alignment contains long segments where the rate doesn't change. In Sub-figure b), the expected value of ν_R is close to 0.5, implying a large number of transitions, and hence generally short segments. ν_R was never observed at such low values in Husmeier (2005) as some unrealistically high rates were included. These drive up ν_R as demonstrated in Sub-figure c), which has the same $\boldsymbol{\rho}_R$ as Sub-figure b) with an unrealistically high rate state added, driving ν_R upwards. See Section 6.10.1 for an explanation of this effect.

6.7.2 Locating multiple behaviours in ν_R when adapting $\boldsymbol{\rho}_R$

Husmeier (2005) fixes the model to a single choice of $\boldsymbol{\rho}_R$. In Figure 6.4, we investigate how the choice of $\boldsymbol{\rho}_R$ affects the posterior of ν_R, the probability of staying in the same rate between adjacent alignment positions. We investigate the alignment of Neisseria described in Section 6.8.1, where the log transition-transversion ratio is fixed at $0.35 = \log_{10} 2.27$, as in Husmeier (2005). Sub-figures a) and b) illustrate that the posterior of ν_R is dependent on the choice of $\boldsymbol{\rho}_R$. The large expected value of ν_R in Sub-figure a) implies that the alignment contains long segments where the rate doesn't change. In Sub-figure b), ν_R peaks at a low value, implying a large number of transitions, and hence on average short segments.

The posterior distribution of ν_R is determined by the choice of $\boldsymbol{\rho}_R$ in a non-obvious fashion. Hence, picking different static choices of $\boldsymbol{\rho}_R$ in order to investigate all behaviours present in the sequence is difficult. In contrast, adapting $\boldsymbol{\rho}_R$ during the simulation will find all behaviours present. Furthermore, ν_R was never found to peak at such low values in Husmeier (2005). This is because some unrealistically high rates were included. These drive up ν_R as demonstrated in Sub-figure c), which has the same $\boldsymbol{\rho}_R$ as Sub-figure b) with an unrealistically high rate state added, driving ν_R upwards. See Section 6.10.1 for an explanation of this effect.

6.7.3 Investigating position specific codon rate variation

In the presence of both large-scale rate heterogeneity as well as codon position specific rate variation, the posterior of ν_R will be multi-modal: one mode representing the large scale behaviour and the other mode reflecting the codon specific rate variation. When a DNA sequence codes for a protein, each triplet in the sequence codes for a single amino acid. However, a change in the third position of the triplet often does not change which amino acid is coded for. Hence, often a mutation in this position has no impact on the function of the protein, and there is a smaller penalty for a mutation in this position. Additionally, even when a mutation in the third codon position changes the resulting amino acid, the biophysical properties of the new amino acid tend to be similar, further implying a lower selective pressure. Given a segment of rate ρ_R, we generate synthetic alignments that exhibit this effect by making each third codon position have rate $\rho_R + c$, where c reflects the strength of the codon specific behaviour. To leave the average rate unchanged, the first and second codon positions have rate $\rho_R - \frac{c}{2}$.

Our synthetic alignment consists of a series of segments of length 150. We first define a segment with $\rho_R = -1.5$, a segment with high rate of $\rho_R = -1$ and a segment with a low rate of $\rho_R = -2$. To investigate how this interacts with topology changes, we also add a region with a different topology to simulate recombination. The segments along the alignment are arranged as: normal, low rate, normal, recombination, normal, high rate, then normal again. We then generate the alignment as outlined in Section 6.6.3. The transition-transversion is inferred using our normal inference scheme.

In Figure 6.5, we investigate the effect of changing c on this synthetic alignment. Sub-figure a) illustrates how the codon positions in each region start to overlap as c increases. In Sub-figure b), we see that the posterior of ν_R becomes multi-modal, as expected. Sub-figure c) represents the posterior for ν_S. We see that as c increases, the posterior for ν_S is not affected - it still correctly peaks at values close to 1, indicating that changes in the topology are unlikely. Sub-figure d) shows how the probability of only having one state varies between the chains. The model correctly predicts that there is only one transition-transversion ratio until c becomes sufficiently large.

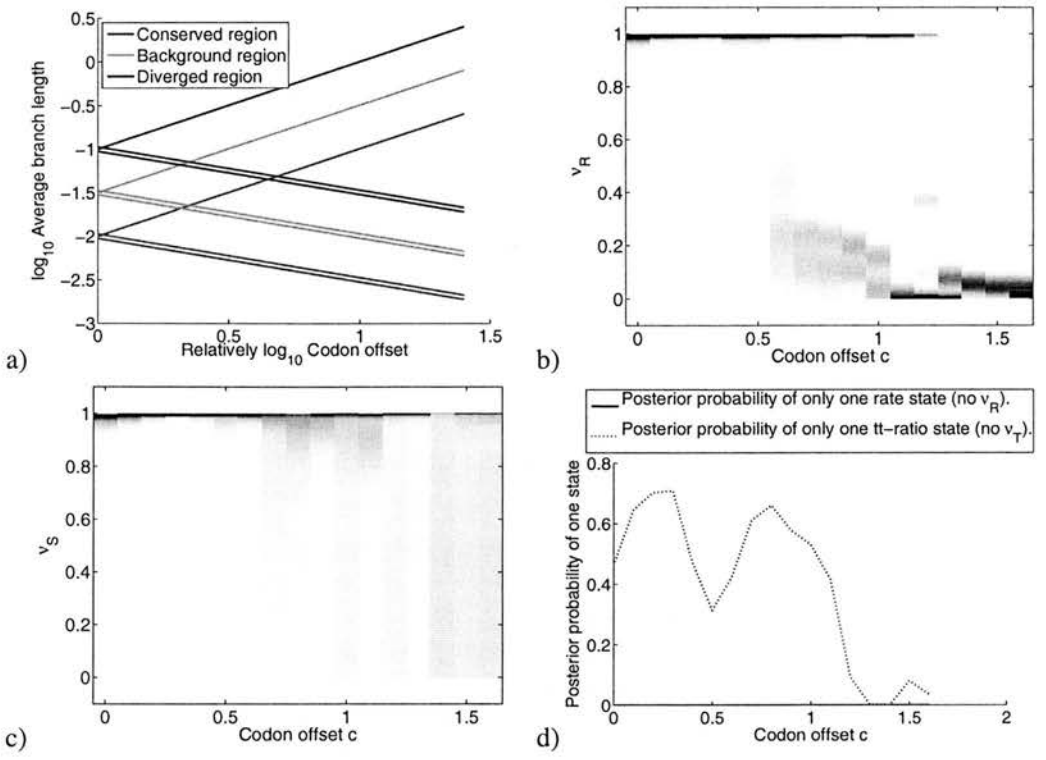


Figure 6.5: Investigating the behaviour of the PRJ-FHMM model in the presence of both large-scale rate heterogeneity and codon position specific rate variation. Here, the log rate of each codon triplet along the sequence is: $\left[\rho_R - \frac{c}{2}, \rho_R - \frac{c}{2}, \rho_R + c\right]$, where ρ_R is the rate of the segment and c reflects the strength of the codon specific behaviour. Sub-figure a) shows the setup of the synthetic study with different shadings indicating different segments. The three lines for each segment indicate the three codon positions. Sub-figure b) shows how the posterior for ν_R varies as we increase c. The horizontal axis represents c while the y-axis displays the posterior distribution of ν_R for that value of c. Darker shadings indicate higher probability. As the codon effect becomes stronger, it starts to dominate the predictions. At $c \approx 0.6$, we can see a multi-modal posterior as the model picks up both behaviours, corresponding to the multiple peaks found in Figures 6.15 and 6.10. Sub-figure c) represents the posterior for ν_S. We see that as c increases, the posterior for ν_S is not affected - it still correctly peaks at values close to 1, indicating that changes in the topology are unlikely. Sub-figure d) shows how the probability of only having one rate and one transition-transversion ratio along the sequence. The model correctly predicts that there is only one transition-transversion ratio until the codon effect is twice as big as the rate difference between the long segments.

6.7.4 Comparison with the MCP of Minin et al. (2005) on a synthetic alignment

We compare the performance of the PRJ-FHMM model and the MCP of Minin et al. (2005) on a simple synthetic multiple-sequence alignment consisting of alternating segments of length 200 at rates -1 and -0.7. The low and high rate segments alternate 6 times, starting and ending with the low mutation rate. In our plots, the thick grey lines will display the underlying true rate compared to the predictions, as seen in Figure 6.7. See Section 6.6.3 for how we generate the synthetic alignment given the rates. This setup, for instance, simulates an alignment consisting of alternating introns and exons.

In Figure 6.6, we investigate the posterior distributions of ν_S, ν_R and ν_T for the synthetic comparison outlined above, where we have set $\lambda_R = 3$. The posterior probabilities of ν_R and ν_T being relevant to modelling the alignment are shown at the top of the appropriate graph. For instance, if the alignment consists of a single rate state, then ν_R is irrelevant – see Section 6.4.5. The posterior distribution of ν_S peaks at values close to 1, correctly indicating that very few topology changes can occur along the alignment. ν_R is relevant as the rate changes along the alignment, and the posterior probability of ν_R being relevant according to the PRJ-FHMM model is 0.988. Additionally, the posterior distribution of ν_R peaks close to the true value of 0.99545, calculated using Equation (6.47) and inserting the number of segments present. The PRJ-FHMM model correctly predicts that it is unlikely that ν_T is relevant. ν_T being irrelevant correctly indicates that the transition-transversion ratio is constant along the alignment.

In Figure 6.7, we investigate how changing the prior on the number of rates affects the predicted number of underlying unique rates and rate segments for the synthetic alignment. This reflects how both methods would be used in practice. We examine the effects of setting the parameters that describe the priors to three different values: A conservative (or most parsimonious) setting, the best setting (where the prior distribution is closest to the posterior) and high variability setting (offset from the best by the difference between best and conservative setting).

The conservative setting is where $\lambda_R = 1$ for the PRJ-FHMM (on average expecting one rate state) and $\lambda_R^B = 1$ for the MCP of Minin et al. (2005), which corresponds to on average expecting one breakpoint. The best setting is where $\lambda_R = 2$ and $\lambda_R^B = 10$. The high variability setting is offset from the best by the difference between the conservative and best settings, hence $\lambda_R = 2 + (2-1) = 3$ and $\lambda_R^B = 10 + (10-1) = 19$.

On average, the MCP of Minin et al. (2005) exhibits a higher error compared to the PRJ-FHMM model, exhibits a noticeably higher variability as seen by the wider confidence intervals, and appears slightly more dependent on the setting of the prior.

In Figure 6.8, we investigate the effect of repeated state visits by using equivalently informative priors on the average number of segments for both models. We re-use the priors on the average number of breakpoints from before: $\lambda_R^B = 1$, $\lambda_R^B = 10$ and $\lambda_R^B = 19$. Using Equa-

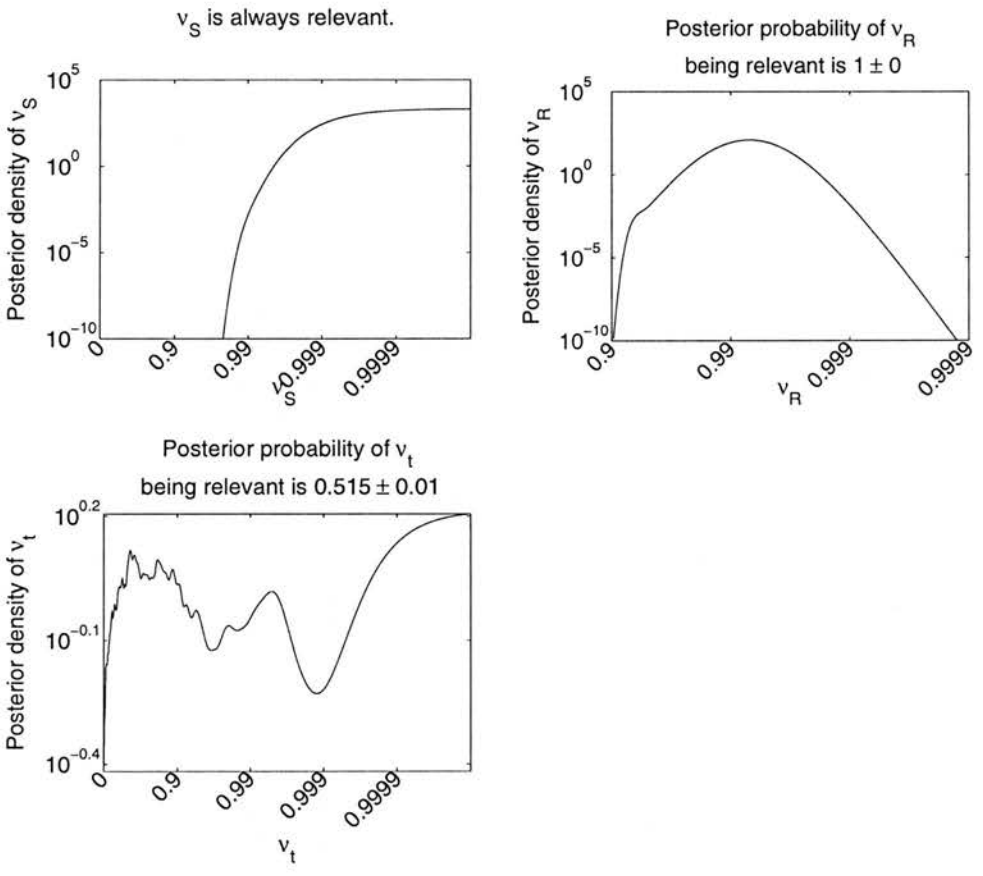


Figure 6.6: The posterior distributions of ν_S, ν_R and ν_T for the synthetic alignment described at the beginning of Section 6.7.4. For each graph, the horizontal axis indicates the value of ν, plotted logarithmically approaching 1, while the vertical axis represents $P(\nu|\mathcal{D})$, the posterior probability of ν, plotted on a logarithmic scale. The posterior probabilities of ν_R and ν_T being relevant to modelling the alignment are shown at the top of the appropriate graph. For instance, if the alignment consists of a single rate state, then ν_R is irrelevant – see Section 6.4.5. The posterior for ν_S peaks at values close to 1, correctly indicating that very few topology changes occur along the alignment. ν_R is relevant as the rate changes along the alignment - the posterior probability of ν_R being relevant according to the PRJ-FHMM model is 0.988. Additionally, the posterior distribution of ν_R peaks close to the true value of 0.99545. The PRJ-FHMM model correct predicts that ν_T is irrelevant, which is correct as the alignment was sampled using a single transition-transversion ratio.

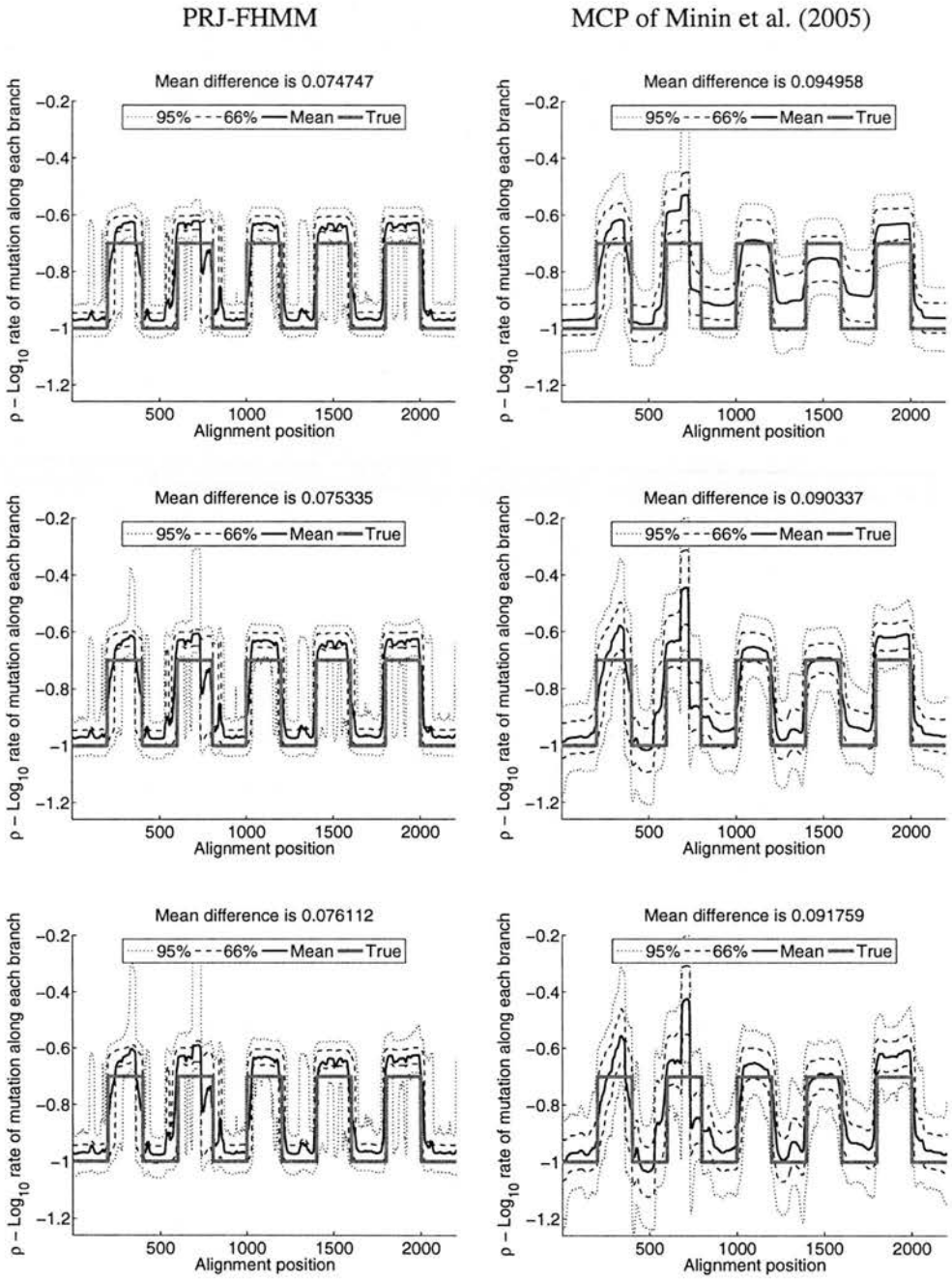


Figure 6.7: A simple synthetic study where the MCP of Minin et al. (2005) has a larger average error and suffers from a slightly stronger dependence on the prior. The alignment consists of repeated alternating segments with rates -1 and -0.7, both 200 columns long. Each graph is set up in an identical fashion to Figure 6.3.

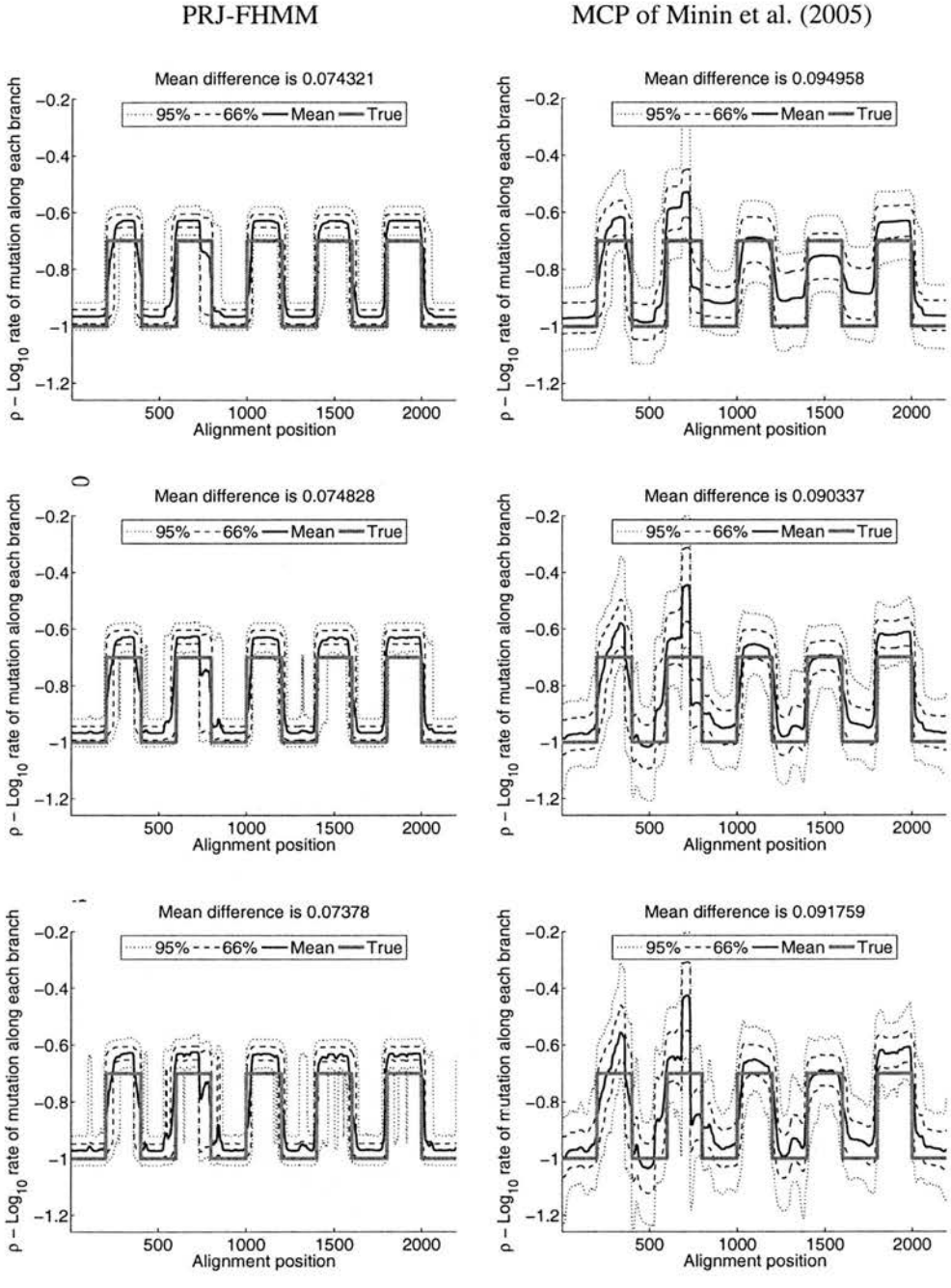


Figure 6.8: We investigate the effect of repeated state visitations by using equivalently informative priors on the average number of rate breakpoints (or rate changes) for both models. We investigate the same values for λ_R^B as in Figure 6.7, which we translate to minimum values for ν_S, ν_R and ν_T for the PRJ-FHMM model using Equation (6.47). See Figure 6.7 for the meaning of the lines and axis in each graph.

tion (6.47) and substituting in $L = 11 * 200 = 2200$, we find that $\nu_R = 0.99955$, $\nu_R = 0.99545$ and $\nu_R = 0.99136$ respectively. We use these values as our lower thresholds on ν_S, ν_R and ν_T, and set $\lambda_R = 1$, where λ_R is the prior on the average number of rate states.

As can be seen from the setting of λ_R^B that makes the prior distribution most closely match the posterior, the MCP cannot as accurately characterise the rate along the sequence as the PRJ-FHMM model. Adding a more informative prior on the average number of segments to the PRJ-FHMM model only slightly reduces the average distance between the mean predicted rate and the true rate, but does produce a cleaner prediction by reducing the 95 percentiles of the credibility intervals of the posterior rate. The average error has not significantly decreased, which implies that the PRJ-FHMM model has already managed to infer ν_S, ν_R and ν_T to some degree of accuracy. The remaining uncertainty was reflected in the uncertain 95 percentiles of the posterior distribution of the rate along the alignment in Figure 6.7.

In Figure 6.9 we investigate the posterior number of rate states found by the PRJ-FHMM model, and the posterior number of rate segments found by both models. Using the same conservative, best and high variability settings from before, the PRJ-FHMM model confidently predicted the correct number of rate states. The number of rate segments predicted by the MCP of Minin et al. (2005) vary strongly as setting of the parameter describing the prior change, making it hard to use the MCP to predict the number of segments present in the alignment. The mode of the PRJ-FHMM model prediction is 17-20 segments, which is an over-estimate. In Sub-figures c) and d) we limit the PRJ-FHMM model to the equivalent of the high variability λ_R^B setting for the MCP. This mapping is done as outlined in Section 6.6.2. Both the posterior number of states and segments are now more highly peaked around their correct values, showing that more informative prior knowledge about the number of segments makes the PRJ-FHMM model more confidently give the correct predictions.

When the PRJ-FHMM model is applied to real alignments, we have the option of directly thresholding ν_R to introduce more prior knowledge about segments, as outlined in Section 6.4.2, or alternatively introducing more informative prior on ν_R by changing the α and β parameters in Equation (6.12). Currently, all values of ν_R are *a priori* equally likely as $\alpha = 1$ and $\beta = 1$. However, we might expect that values of ν_R closer to 1, corresponding to longer segments, to be more likely, which can be specified by adjusted α and β.

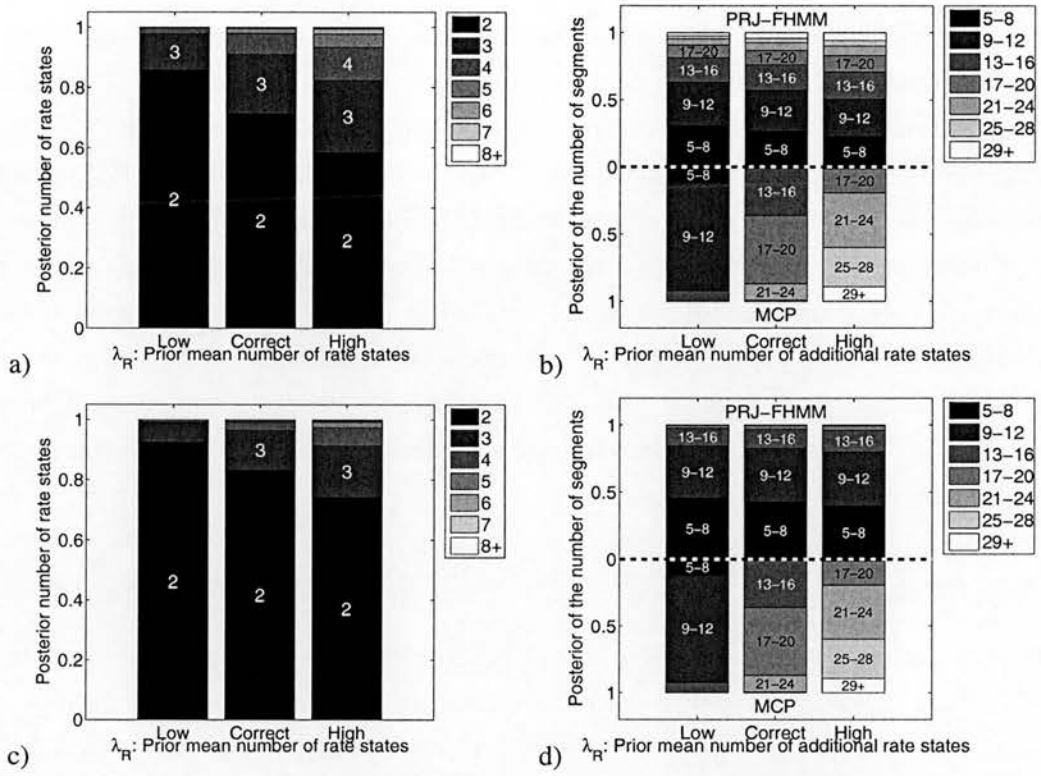


Figure 6.9: The posterior number of rate states found by the PRJ-FHMM model, and the posterior distribution of the number of rate segments for both models on the synthetic alignment described in Section 6.7.4. Sub-figure a) shows for the PRJ-FHMM model how the posterior number of rate states change when using the conservative, best and higher variability priors from before (labelled as low, correct and high priors respectively). The horizontal axis ranges across these priors, where for each prior the posterior distribution over number of states is shown vertically, summing to one. For all priors tested, the PRJ-FHMM model confidently and correctly predicted that there are two rate states. Sub-figure b) compares the posterior number of rate segments between the models, where bar-graphs that point upwards represent the posterior number of segments for the PRJ-FHMM model, while the bar-graph that point downwards represents the posterior number of segments for the MCP of Minin et al. (2005). The mode of the PRJ-FHMM models predictions is at 17-20 segments, which is an over-estimate. In Sub-figures c) and d), we instead of setting λ_R^B, we lower-bound ν_S, ν_R and ν_T from PRJ-FHMM model to the equivalent of λ_R^B from the MCP of Minin et al. (2005).

6.8 Application to real DNA sequence alignments

6.8.1 Neisseria

One of the first indications for sporadic recombination was found in the bacterial genus *Neisseria* (Maynard Smith, 1992). We selected the four strains *Neisseria gonorrhoeae* (X64860), *Neisseria meningitidis* (X64866), *Neisseria cinera* (X64869) and *Neisseria mucosa* (X64873), where GenBank/EML accession numbers are shown in brackets. Zhou and Spratt (1992) found two anomalous, or more diverged regions in the DNA alignment, which occur at positions $t = 1 - 202$ and $t = 507 - 538$ (Note that Zhou and Spratt, 1992 used a different labelling scheme, with the first nucleotide at $t = 296$, and the last one at $t = 1082$.) In the rest of the alignment, *N.meningitidis* clusters with *N.gonorrhoeae* (defined as topology $H_{S,t} = \rho_{R,1}$ in our HMM), while between $t = 1$ and $t = 202$, they found that it is grouped with *N.cinera* (defined as state $H_{S,t} = 3$). Zhou and Spratt (1992) suggested that the region $t = 507 - 538$ is the result of rate variation.

6.8.2 Maize

We investigate an alignment of the gene family of maize actin genes where gene conversion has been found to occur – a process similar to recombination. This process occurs in multi-gene families, where a DNA subsequence of one gene can be replaced by the DNA subsequence from another. Indication of gene conversion between two pairs of maize actin genes has been reported in Moniz de Sa and Drouin (1996). We use the same alignment and specification of the topologies as Husmeier and McGuire (2003).

6.8.3 HIV-1 KAL-153

In 1996, a recombinant HIV-1 strain termed KAL153 caused an epidemic outbreak of AIDS infection among intravenous drug users around Kaliningrad, Russia. Liitsola et al. (1998) identified KAL153 as a recombinant of subtypes A and B. We use a whole genome sequence alignment of KAL153 with three consensus sequences of subtypes A, B and F from the Los Alamos HIV Sequence Database. We used the same sequence alignment as in Suchard et al. (2003).

6.8.4 Simulation settings and an empirical comparison with the MCP of Minin et al. (2005)

For an empirical comparison between the proposed phylogenetic FHMM and the dual MCP of Minin et al. (2005), we map $\lambda_R^B = 2\lambda_R - 1$. This assumes that on average every extra rate state causes two extra breakpoints and that when $\lambda_R^B = 1$, both models are set to their most

conservative prior distributions. We investigate the stability of the methods by investigating the settings: $\lambda_R \in \{1,\ldots,5\}$, which is thus mapped to $\lambda_R^B \in \{1,3,5,7,9\}$ for the MCP of Minin et al. (2005). In order to easily compare our inference scheme to that used by the MCP, we use their Gaussian priors on $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$, as described in Section 6.4.2. We investigate the effect of using the uniform and the even number order statistics prior in Section 6.9.4.

This leaves us to specify λ_S^B from Equation (6.45), the mean number of topology breakpoints expected. We follow their recommendations and set $\lambda_S^B = \sqrt{2}$. In contrast, the PRJ-FHMM model infers ν_S, the equivalent parameter.

6.9 Results on real sequence alignments

6.9.1 Segmenting the alignment of Neisseria DNA sequences

6.9.1.1 The posterior probability distributions of ν_S, ν_R and ν_T

In Figure 6.10, we explore the posterior distributions of ν_S, ν_R and ν_T for the alignment of Neisseria DNA sequences described in Section 6.8.1. This initial exploration was performed with $\lambda_R = 3$, the middle of the range of investigated priors. The posterior probabilities of ν_R and ν_T being relevant to modelling the alignment are shown at the top of the appropriate graph. ν_R is relevant with probability 0.982 ± 0.003. Hence, it is highly likely that the alignment exhibits rate heterogeneity – see Section 6.4.5. In contrast, the posterior probability of ν_T begin relevant is only 0.41 ± 0.005, indicating the transition-transversion ratio probably does not vary along the alignment. ν_S is always relevant, as $\boldsymbol{\rho}_S$ and k_S are fixed by the construction of our model.

Notice that the posterior distribution of ν_S has only a single mode, while the posterior distribution of ν_R has multiple modes. The multi-modality is presumably due to a codon position specific rate variation. To test this conjecture, we synthetically generated a set of DNA sequence alignments with different trade-offs between codon-specific and region-specific rate variation – see Section 6.7.3. The results of this investigation on a synthetic alignment suggests that the bimodality observed in the distribution of $P(\nu_R)$ on Neisseria could, indeed, result from an interplay of these two effects. This implies that the peak around $\nu_R = 0.9$ could be due to the codon effect. We are mainly interested in region-specific rate heterogeneity, owing to its confounding effect on the detection of recombination (Husmeier, 2005), and its increasing relevance in functional genomics (Nimrod et al., 2005; Siepel and Haussler, 2004). For that reason we focus on the peak representing promising long scale behaviour at around $\nu_R = 0.995$, by setting $C_R^{\min} = 0.667$, the minimum between the codon and region effect peaks.

ν_T also exhibits multiple modes, indicating that there might be multiple different behaviours in the original alignment. However, the low posterior probability of ν_T being relevant indicates that the transition-transversion ratio is probably invariant along the alignment. Again,

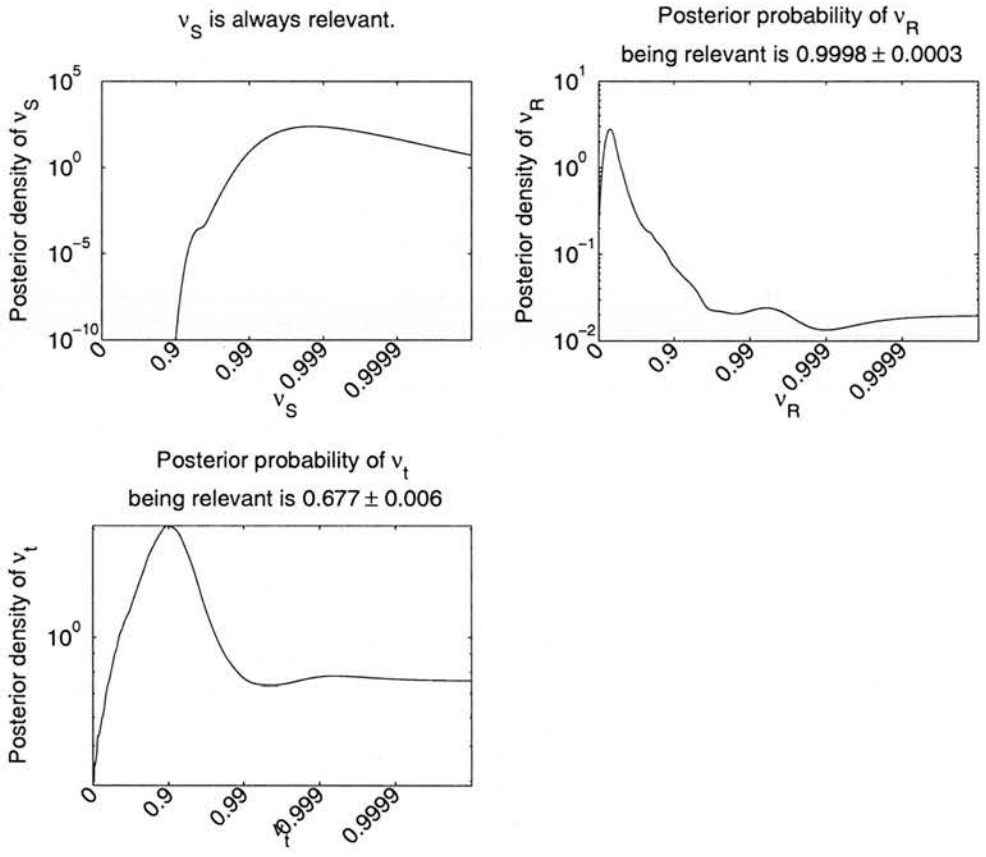


Figure 6.10: The posterior distributions of ν_S, ν_R and ν_T for the alignment of Neisseria DNA sequences described in Section 6.8.1. This figure is laid out in an identical fashion to Figure 6.15. The posterior probability of ν_R being relevant for modelling the alignment is 0.982 ± 0.003 (indicated at top of graph), where ν_R is irrelevant if there is only a single rate state ($k_R = 1$). The posterior probability of ν_T begin relevant is 0.41 ± 0.005, indicating the transition-transversion ratio probably does not vary along the alignment.

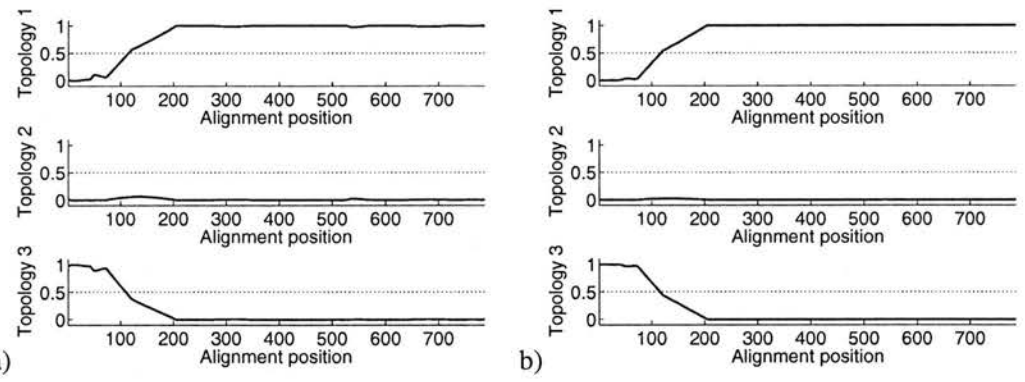


Figure 6.11: The posterior distribution of the phylogenetic tree topology along the alignment of Neisseria DNA sequences. Sub-figure a) shows the predictions for the PRJ-FHMM while Sub-figure b) shows the predictions for the MCP of Minin et al. (2005). The x-axis represents the alignment position, the y-axis the probability, and each sub-plot indicates the posterior probability of each possible topology. Topology 1 : [(*N.gonorrhoeae*, *N.meningitidis*), (*N.cinera*, *N.mucosa*)]; Topology 2 : [(*N.gonorrhoeae*, *N.cinera*), (*N.meningitidis*, *N.mucosa*)]; Topology 3: [(*N.gonorrhoeae*, *N.mucosa*), (*N.cinera*, *N.meningitidis*)]. Zhou and Spratt (1992) predicted a breakpoint at position 202, while the phylogenetic FHMM predicts it to lie in the region 80-200 - this is the region where the posterior probability of the recombinant tree topology gradually decreases from 1 to 0.

we are interested in long scale variation, and set $C_T^{\min} = 0.995$.

This experiment was then rerun, and $C_R^{\min}$ and $C_T^{\min}$ were set to the values of ν_R and ν_T with the lowest posterior probability. This was repeated until $C_R^{\min}$ and $C_T^{\max}$ were located at the minimums of ν_R and ν_T, which occurred when $C_R^{\min} = 0.977$ and $C_T^{\max} = 0.995$.

In all cases, the posterior distribution of ν_S peaked at values close to 1, indicating that *a posteri* there are only a few topology changes in the alignment.

6.9.1.2 Posterior distributions of the phylogenetic tree topology and rate

In Figure 6.11, we investigate the posterior distribution of the phylogenetic tree topology for the PRJ-FHMM and the MCP of Minin et al. (2005). The predictions are in good agreement with those of Zhou and Spratt (1992) – see the caption of the figure for details. We display the predictions for $\lambda_R = 1$ and $\lambda_R^B = 1$ – the predictions appeared stable for the ranges of priors on the rates we tested: $\lambda_R \in \{1,\ldots,5\}$ and the corresponding $\lambda_R^B \in \{1,3,5,7,9\}$

Figure 6.12 compares the predicted rate along the alignment for the PRJ-FHMM and the MCP of Minin et al. (2005). We display the results for the extremities of the prior range. The PRJ-FHMM consistently finds that regions 1-75 and 425-530 are more diverged, with some minor divergence around 345-385. In contrast, the MCP is strongly dependent on the prior

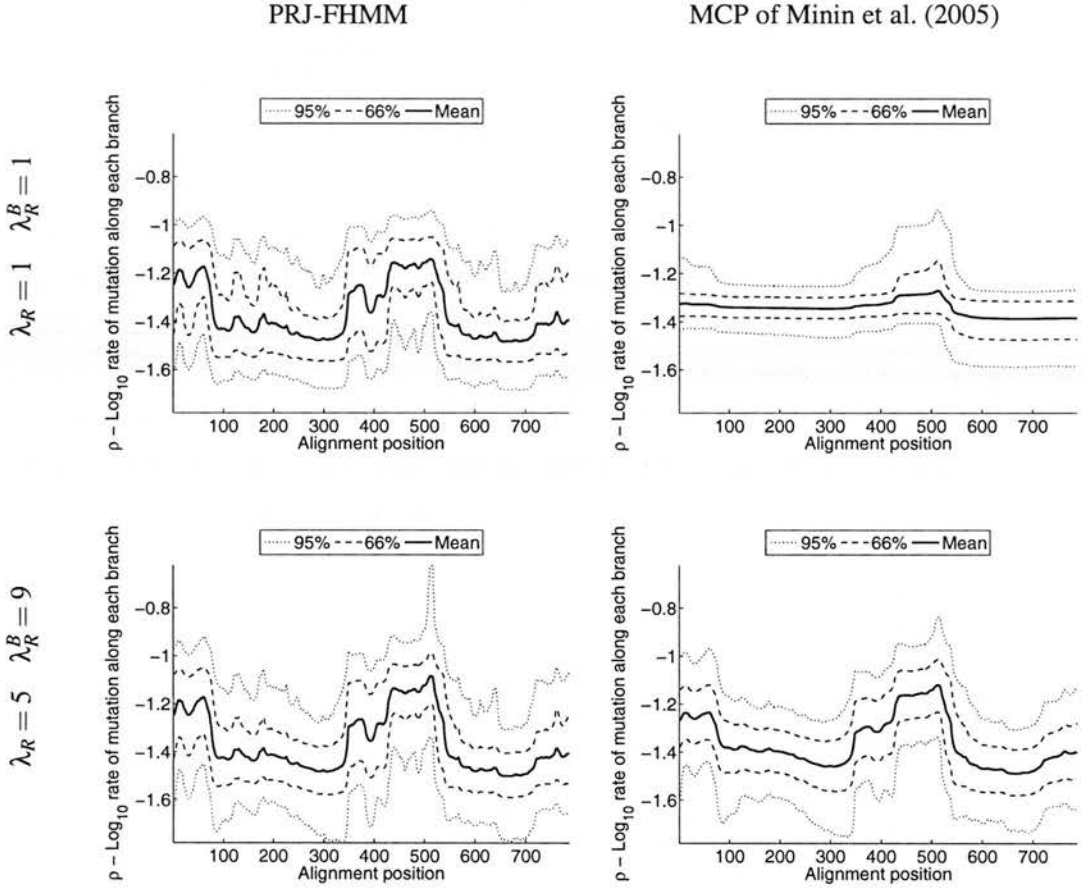


Figure 6.12: The credibility intervals of the posterior rate distribution along the alignment of Neisseria DNA sequences. In each panel, the x-axis indicates the alignment position and the y-axis indicates the log rate. The black line represents the mean predicted rate, while the dashed and dotted lines represent the 66th and 95th credibility intervals. The panels on the left represent the PRJ-FHMM, while the panels on the right represent the MCP. The top and bottom panels compare the lowest and highest values of λ_R (hence also λ_R^B, due to the mapping described in Section 6.6.2) that were investigated.

as the detection of rate variation at positions 1-75 is found only for specific settings of the prior. Zhou and Spratt (1992) found two anomalous regions present in the alignment: 1-202 and 507-538. They suggested that 1-202 is the result of recombination, while unsure of the origin of the anomalous region 507-538. Both the PRJ-FHMM and the MCP agree with Zhou and Spratt (1992) that no topology change occurs around the region 507-538, but identify the anomalous region as part of a larger region of rate variation, namely 425-530. The PRJ-FHMM and MCP consistently identified that a recombination event occurs towards the beginning of the sequence. However, only the PRJ-FHMM consistently identified the region 1-75 as being more diverged. In contrast, the MCP predictions were dependent on the setting of λ_R^B.

Figure 6.13 shows comparisons of the predicted numbers of rate states and segments by the PRJ-FHMM and the MCP of Minin et al. (2005), and their dependence on the prior. The PRJ-FHMM gives a stable prediction of the number of different rate states present, which neither the MCP nor the model of Husmeier (2005) can estimate. For $\lambda_R \in \{1,\ldots,5\}$, the PRJ-FHMM predicts that there are two to three rate states that occur repeatedly along the alignment. Sub-figure b) shows the predicted number of segments with the PRJ-FHMM and MCP shown in the top and bottom panels, respectively. The MCP cannot predict the number of states present, and furthermore is sufficiently sensitive to changes in λ_R that predicting the number of segments present is extremely difficult due to our lack of knowledge about how λ_R is set. In contrast, the predictions of the PRJ-FHMM are stable – a consequence of the fact that in the PRJ-FHMM, we infer the distribution over ν_R from the DNA sequence alignment. This produces stable predictions which indicate uncertainty over the true number of segments present. Sub-figures c) and d) show the effect of increasing prior knowledge on the PRJ-FHMM, where increasing the amount of prior knowledge makes the predictions of the number of segments found more in line with our expectations.

6.9.1.3 Posterior distribution of the transition-transversion ratio.

Figure 6.14 shows the credibility intervals of the posterior distribution of the transition-transversion ratio along the alignment for both models. The predicted transition-transversion ratio along the alignment is very stable for the PRJ-FHMM model and independent of the prior on rate states. In contrast, the predictions of the MCP of Minin et al. (2005) show slightly more variation, presumably due to its tying together of the rate and transition-transversion ratio into the same segment. As shown in the Figure 6.10, the most likely explanation for the transition-transversion ratio found by the PRJ-FHMM model is that there is only a single log transition-transversion ratio for the whole alignment, which is reflected in the predicted posterior transition-transversion ratio.

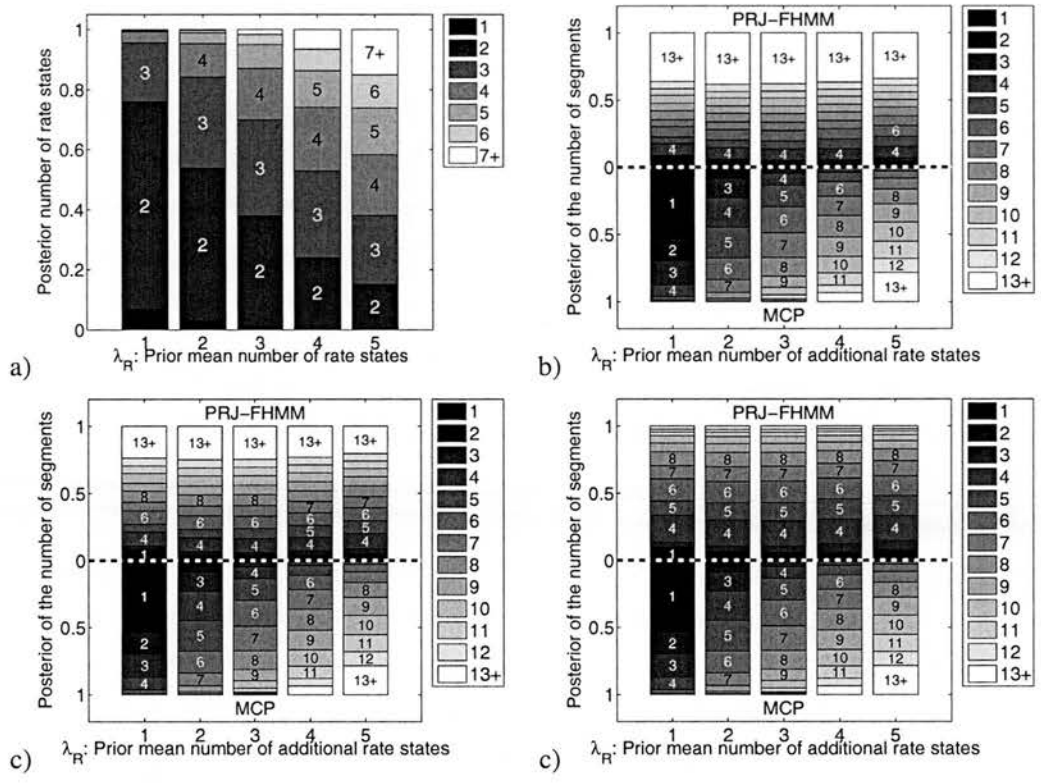


Figure 6.13: Comparisons of the predicted numbers of rate states and segments present on the alignment of Neisseria DNA sequences by our PRJ-FHMM and the MCP of Minin et al. (2005). In all cases, the horizontal axis represents λ_R, the expected mean number of rate states, mapped to the MCP as described in Section 6.6.2. Sub-figure a) shows the predicted number of states. The PRJ-FHMM gives a stable prediction of the number of different rate states present, which neither the MCP nor the model of Husmeier (2005) can estimate. For $\lambda_R \in \{1,\ldots,5\}$, the PRJ-FHMM predicts that there are two to three rate states that occur repeatedly along the alignment. Sub-figure b) shows the predicted number of segments with PRJ-FHMM and MCP (depicted in the top and bottom panel, respectively). The MCP is sufficiently sensitive to changes in λ_R that predicting the number of segments present is highly dependent on prior knowledge. The segment predictions for the PRJ-FHMM are uncertain, with 1, 4 or 6 segments most likely. Sub-figures c) and d) show the effect of increasing prior knowledge on the PRJ-FHMM, where in Sub-figure c) ν_R is constrained such that $0.98 < \nu_R$, equivalent to expecting that segments are not on average shorter than 50 base pairs. Sub-figure d) shows the result of constraining ν_R such that the average number of segments is at most 9. This is equivalent to the distribution over segments when $\lambda_R = 5$ for the MCP.

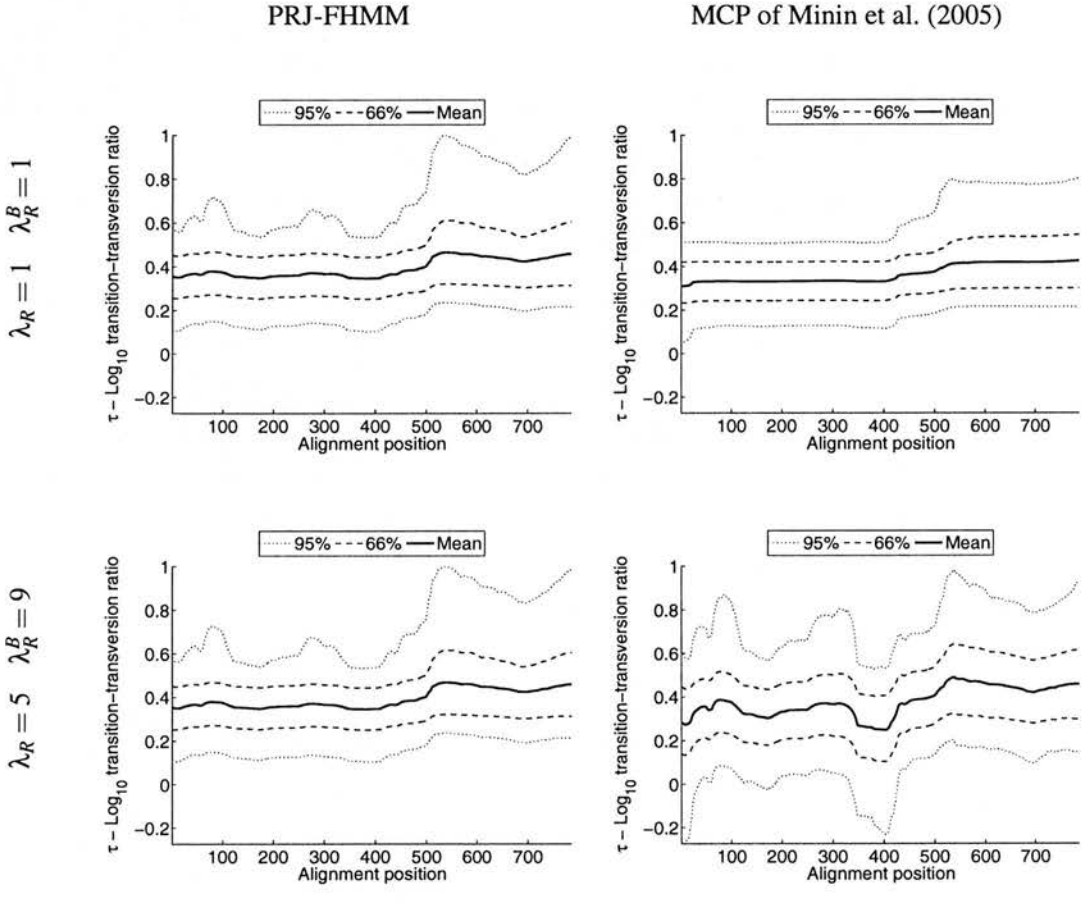


Figure 6.14: The credibility intervals of the posterior distribution of the log transition-transversion ratio along the alignment of Neisseria DNA sequences. This figure is laid out identically to Figure 6.12, expect that the transition-transversion ratio is plotted, not the rate. The predictions of the PRJ-FHMM model are more stable to changes in the rate prior, while the predictions of the MCP of Minin et al. (2005) are slightly more variable.

6.9.2 Segmenting the alignment of maize DNA sequences

6.9.2.1 The posterior probability distributions of ν_S, ν_R and ν_T

In Figure 6.15, we examine the posteriors distributions for ν_S, ν_R and ν_T on the alignment of maize DNA sequences described in Section 6.8.2. Again, for our initial explorations of the alignment, we set $\lambda_R = 3$, the middle of the range of settings investigated. The posterior probability for ν_R being relevant to modelling the alignment is 0.903 ± 0.005, hence, it is highly likely that the alignment exhibits rate heterogeneity (see Section 6.4.5). In contrast to ν_R, the posterior probability of ν_T being relevant is 0.18 ± 0.01, indicating that it is likely that the transition-transversion ratio does not vary along the alignment.

The posterior distribution of ν_R contains multiple modes. As with the alignment of Neisseria in Section 6.9.1, the peak around 0.5 probably corresponds to the codon effect – see Section 6.7.3 for our synthetic experiments that investigate this effect. The peak representing promising long scale behaviour is at around $\nu_R = 0.995$. Like with the alignment of Neisseria DNA sequence, we focus on the peak representing promising long scale behaviour at around $\nu_R = 0.995$, by setting $C_R^{\min} = 0.963$, the minimum between the codon and region effect peaks.

ν_T also exhibits multiple modes, indicating that there might be multiple different behaviours in the original alignment. The low posterior probability of ν_T being relevant indicates that the transition-transversion ratio is probably invariant along the alignment. Again, we are interested in long scale variation, and set $C_T^{\min} = 0.992$.

This experiment was then rerun, and $C_R^{\min}$ and $C_T^{\min}$ were set to the values of ν_R and ν_T with the lowest posterior probability. This was repeated until $C_R^{\min}$ and $C_T^{\max}$ were located at the minimums of ν_R and ν_T, which occurred when $C_R^{\min} = 0.677$ and $C_T^{\max} = 0.994$. As with our analysis of the alignment of Neisseria DNA sequences, the highly peaked distribution of ν_S shows that the model predicts that there are only few topology changes in the alignment.

6.9.2.2 Posterior distributions of the phylogenetic tree topology and rate

In Figure 6.16, we investigate the posterior distribution of the phylogenetic topology along the alignment of maize DNA sequences, where $\lambda_R = 1$ and $\lambda_R^B = 1$. The predictions appeared stable for the ranges of priors on the rates we tested. The posterior topology distributions are very similar between the models, as both detect the topology changing somewhere between alignment positions 700 to 980 - this is where the posterior probability of topology 1 changes from 1 to 0. This change in topology is evidence that gene-conversion occurred along these sequences.

Figure 6.17 compares the credibility intervals of the posterior rate distribution for both models, for the extremities of the prior range. The area of low selective pressure (or high rate) detected between positions 370 and 465, might be an intron that was inadvertently included in

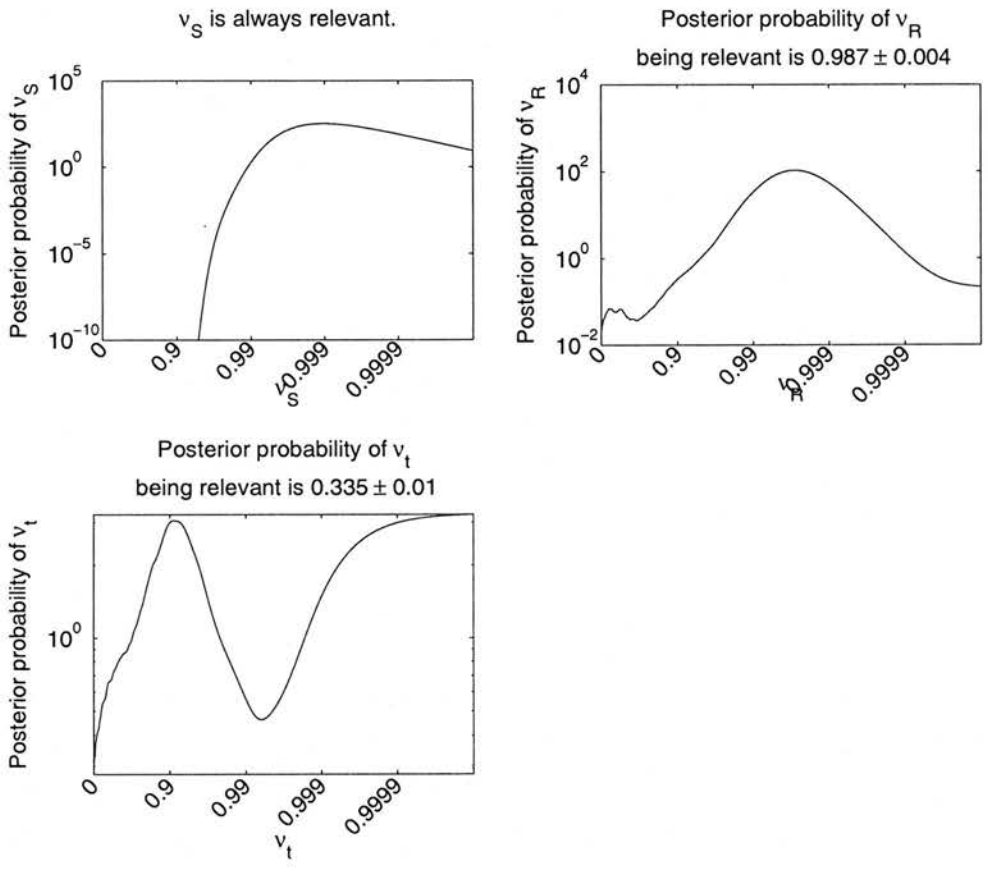


Figure 6.15: The posterior distributions of ν_S, ν_R and ν_T for the alignment of maize DNA sequences described in Section 6.8.2. This is laid out in an identical fashion to Figure 6.10. The posterior probability of ν_R being relevant for modelling the alignment is 0.903 ± 0.005 (indicated at top of graph), where ν_R is irrelevant if there is only a single rate state ($k_R = 1$), while the posterior probability of ν_T being relevant is only 0.178 ± 0.01, indicating that it is highly likely that the transition-transversion ratio is constant along the alignment. Both ν_R and ν_T exhibit multiple modes, indicating that there might be multiple different behaviours in the original alignment.

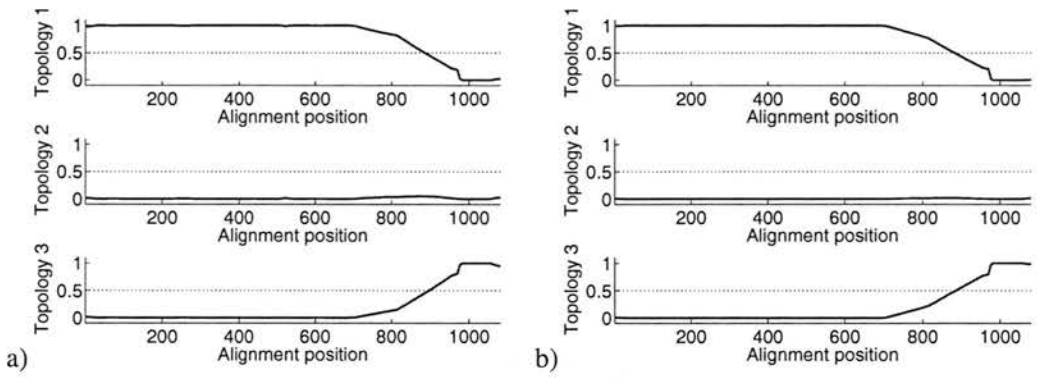


Figure 6.16: The posterior distribution of the phylogenetic tree topology along the alignment of maize DNA sequences. Sub-figure a) shows the posterior distribution of the topology along the alignment for the PRJ-FHMM model, while Sub-figure b) shows the corresponding posterior distribution for the MCP of Minin et al. (2005). In both cases, the x-axis represents the alignment position, and each sub-plot indicates the posterior probability of each possible topology. The posterior topology distributions are very similar between the models, as both detect the topology changing somewhere between alignment positions 700 to 980 - this is where the posterior probability of topology 1 changes from 1 to 0.

the alignment, as indicated by its significantly higher rate.

In Figure 6.18, we investigate the number of predicted rate states for the PRJ-FHMM model, and posterior number of rate segments for both models, and their dependence on λ_R. For $\lambda_R \in \{2,3,4\}$, the PRJ-FHMM model predicts that there are 3 rate states in the alignment, and it is most likely that there are three segments along the segment. We also investigate the effect of more informative priors on ν_R.

6.9.2.3 Posterior distribution of the transition-transversion ratio.

In Figure 6.19, we investigate the posterior distribution of the log transition-transversion ratio along the alignment. Comparing the columns on the left, we see that the PRJ-FHMM model gives stable predictions independent of the rate prior, while the predictions of the MCP of Minin et al. (2005) are slightly more variable. This is to be expected, as the PRJ-FHMM decouples predictions of the rate from those of the transition-transversion ratio. As shown in Figure 6.15, the most likely model for the transition-transversion ratio by the PRJ-FHMM model is that there is only a single transition-transversion ratio for the whole alignment, which is reflected in the predicted posterior transition-transversion ratio.

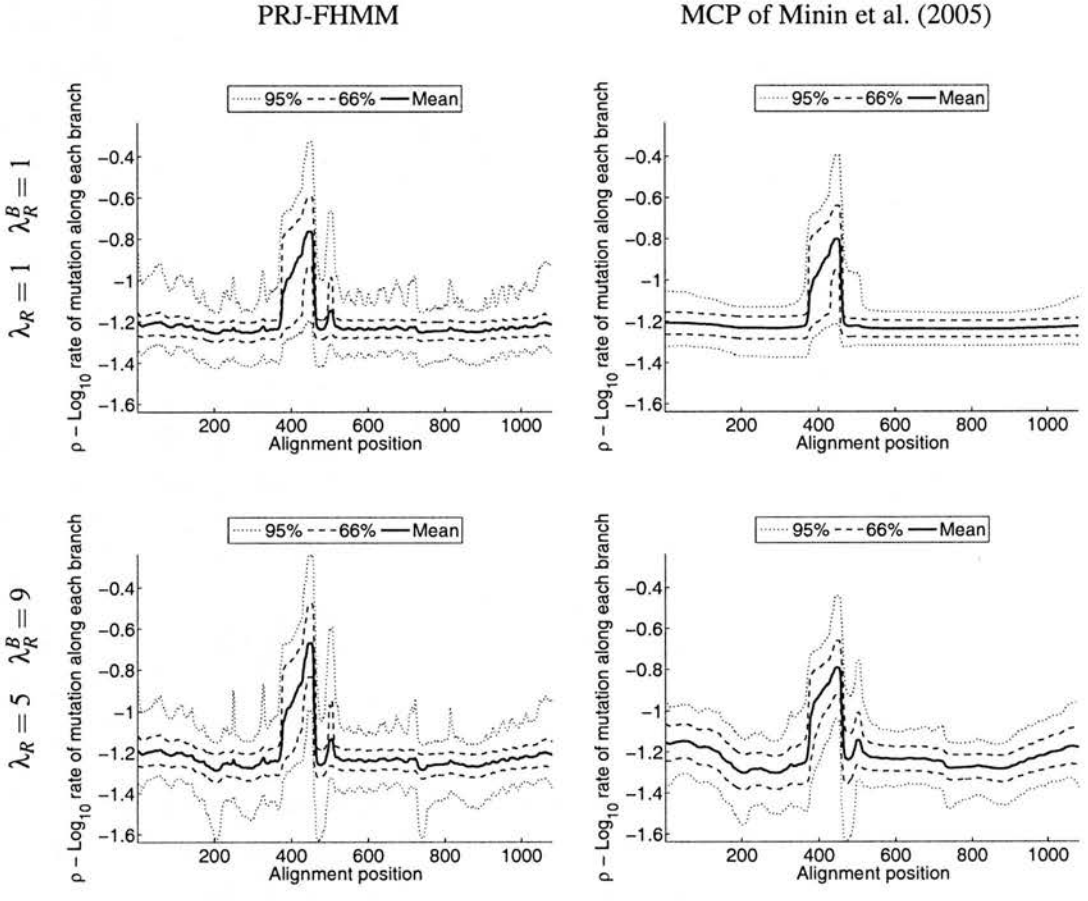


Figure 6.17: The credibility intervals of the posterior rate distribution along the alignment of maize DNA sequences for the PRJ-FHMM and the MCP of Minin et al. (2005), laid out in an identical fashion to Figure 6.12. Both models detect an area of low selective pressure (or high rate) between positions 370 and 465, which might be an inadvertently included intron, as indicated by its significantly higher rate.

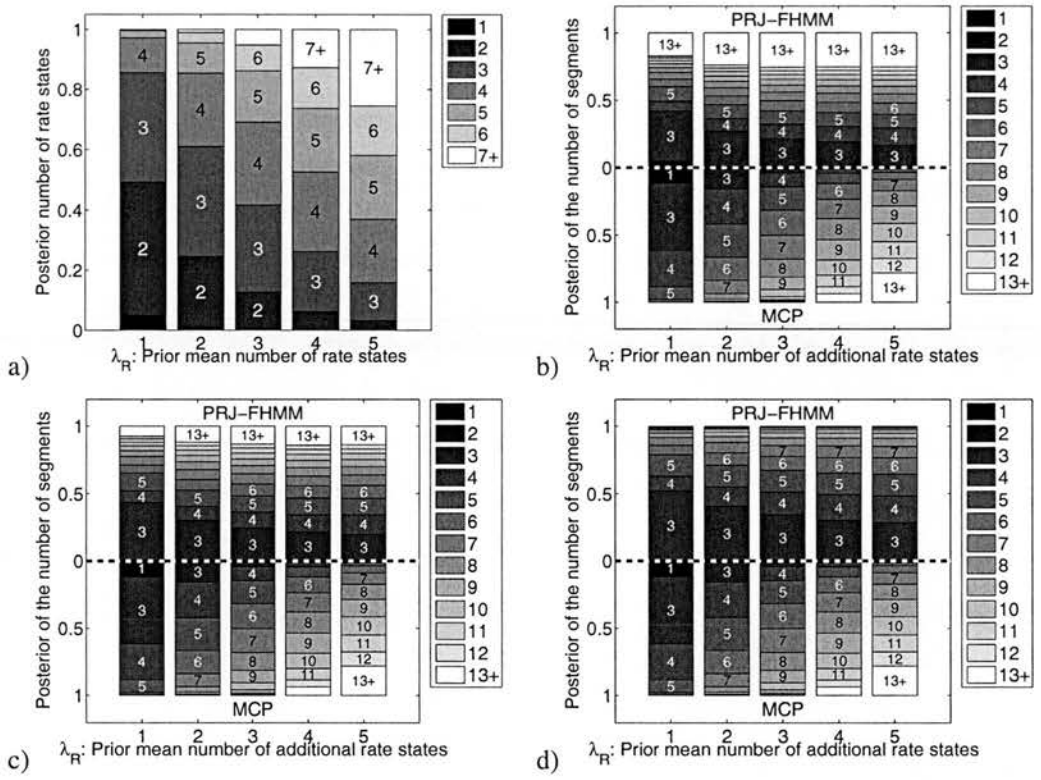


Figure 6.18: Comparisons of the predicted numbers of rate states and segments present on the alignment of maize DNA sequences by our PRJ-FHMM and the MCP of Minin et al. (2005). This is laid out in an identical fashion to Figure 6.13. For $\lambda_R \in \{2,3,4\}$, the PRJ-FHMM model predicts that there are 3 rate states in the alignment. It is most likely that there are three segments according to the PRJ-FHMM model. The MCP is sufficiently sensitive to changes in λ_R that predicting the number of segments present is highly dependent on prior knowledge. In Sub-figure c) ν_R is constrained such that $0.98 < \nu_R$, equivalent to expecting that segments are not on average shorter than 50 base pairs. Sub-figure d) shows the result of constraining ν_R such that the average number of segments is at most 9.

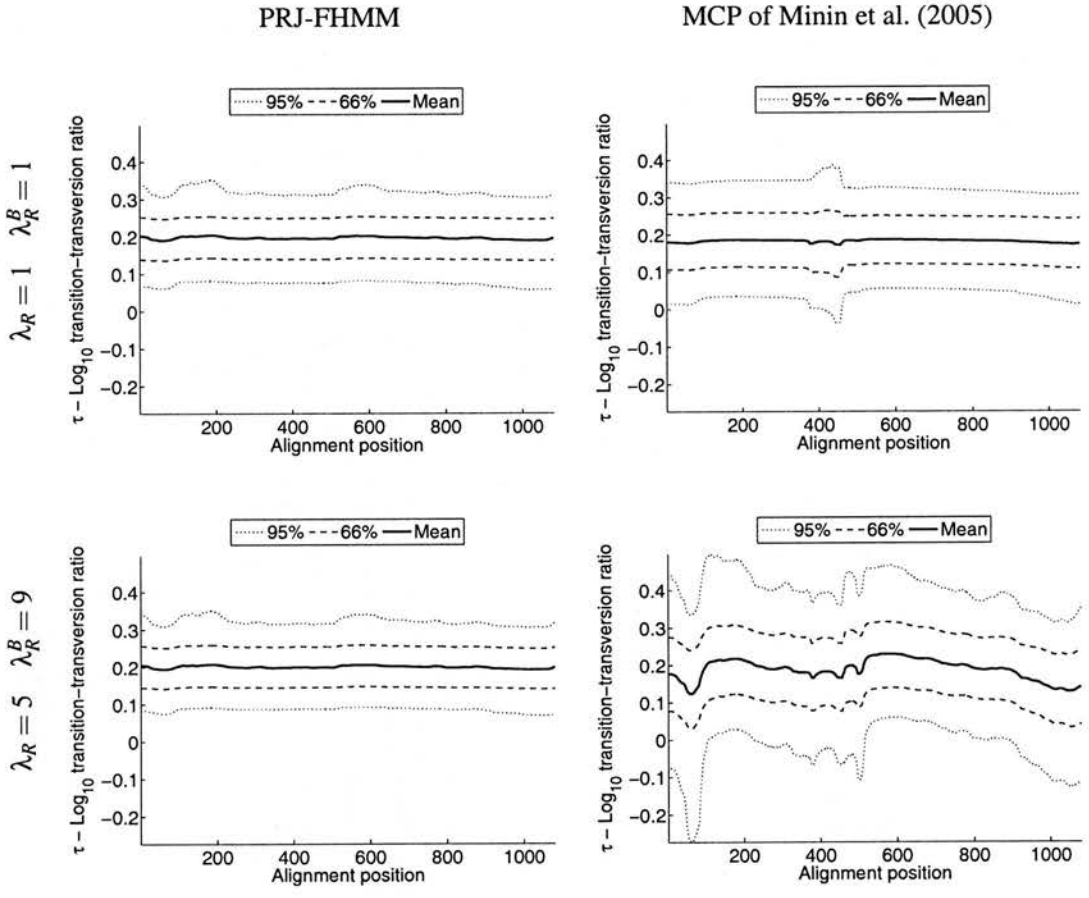


Figure 6.19: The credibility intervals of the posterior distribution of the log transition-transversion ratio along the alignment of maize DNA sequences, laid out as in Figure 6.14.

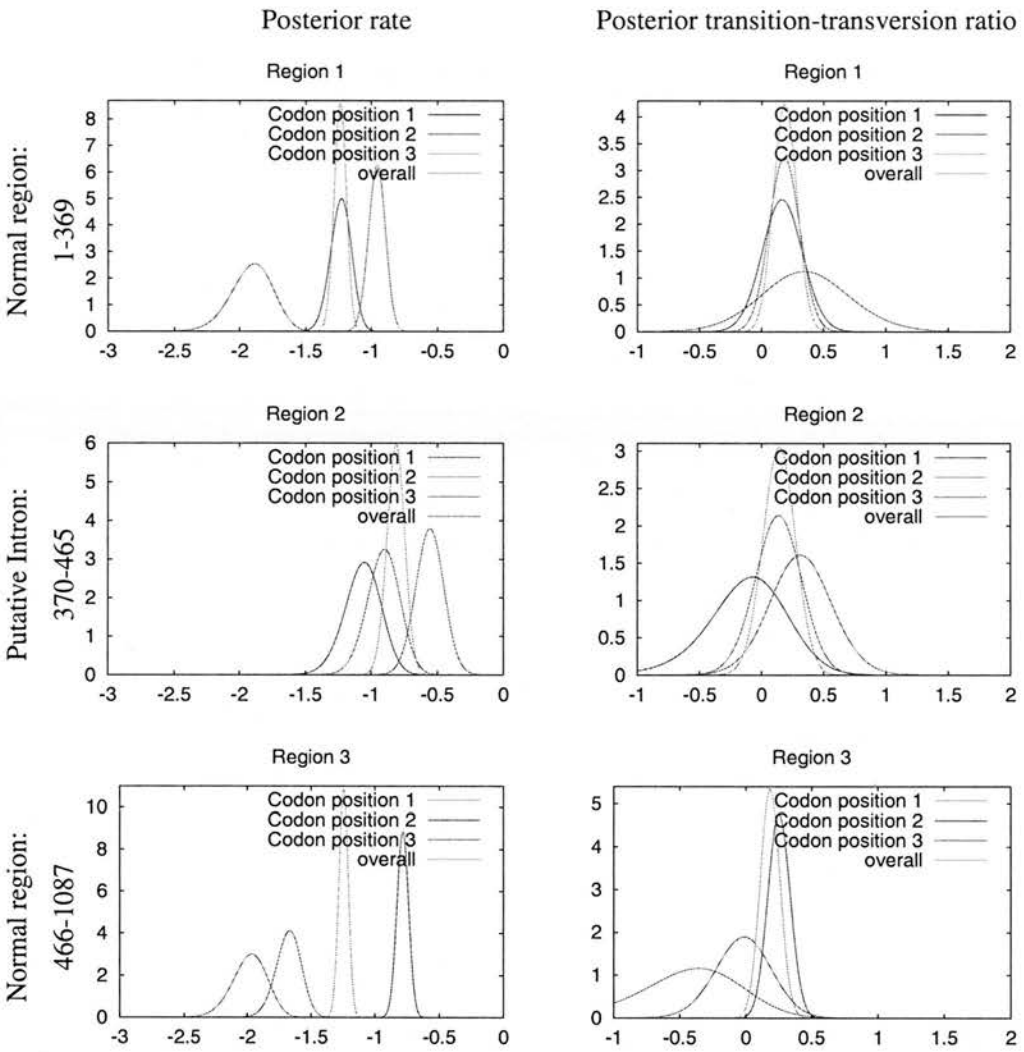


Figure 6.20: Investigating why the codon effect is significantly smaller in the alignment of Maize DNA sequences. The panels on the left show the posterior rate for each codon position in that segment, while the panels on the right show the corresponding codon position posterior distributions of the transition-transversion ratio. The middle panels show the per codon position posterior distributions for the area of high rate seen in Figure 6.17, while the panels at the top and bottom show the corresponding per codon position distribution for the region before and afterwards.

6.9.2.4 Investigating the codon effect on Maize

Figure 6.20 investigates why the codon effect is significantly smaller along the alignment than along the alignment of Neisseria DNA sequences. The graphs indicate that the region 370-465 does not exhibit a codon effect, while the rest of the sequence does. This explains the smaller size of the codon effect peak in the posterior distribution of ν_R seen in Figure 6.15 as there is no codon effect in this middle region. This also provides additional evidence that the spike in the rate is caused by an inadvertently included intron, as introns would not generally exhibit a codon effect.

The posterior distributions of the transition-transversion ratio for each codon position does not provide strong evidence for a codon position specific transition-transversion ratio effect.

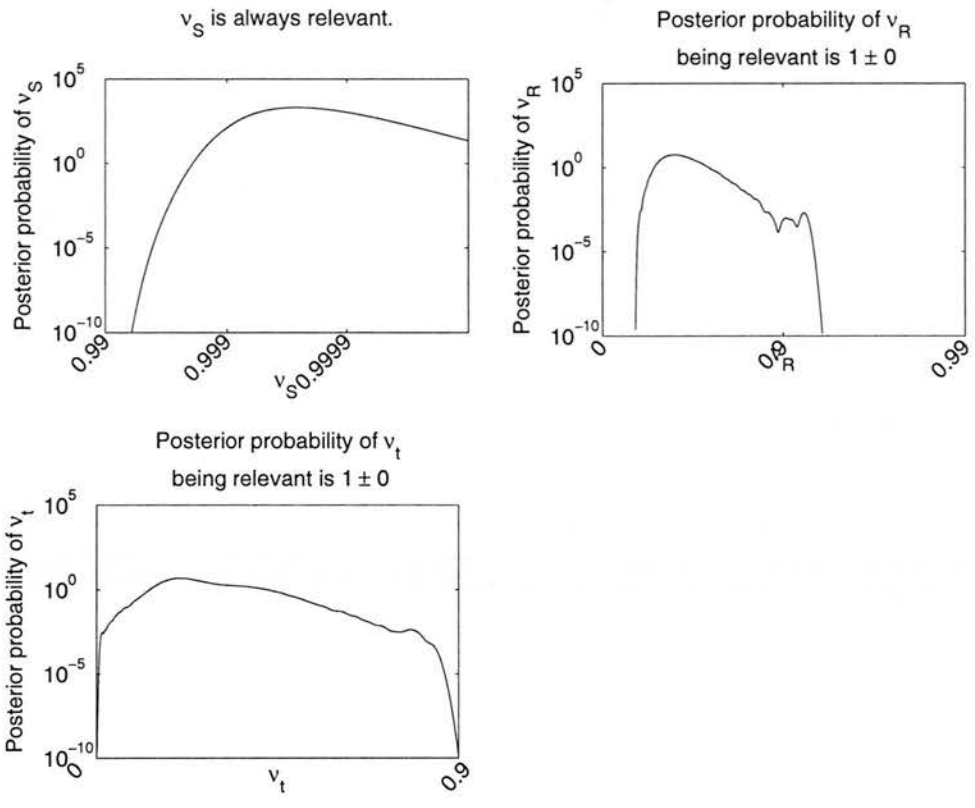


Figure 6.21: The posterior distributions of ν_S, ν_R and ν_T for the alignment of HIV-1 DNA sequences described in Section 6.8.3. The expected value of ν_S is 0.995, clearly shows that the model predicts only a few topology changes. The posterior distribution of ν_R indicates more changes in the rate than expected - see Section 6.10.4 for some possible explanations for this. We instead threshold ν_R and ν_T such that $0.999 < \nu_R$ and $\nu_T < 0.999$, indicating that there are not more than 9 rate or transition-transversion ratio breakpoints in the alignment, which is equivalent to the prior used by Minin et al. (2005). ν_S is correctly inferred by the model.

6.9.3 Segmenting the alignment of HIV-1 DNA sequences

6.9.3.1 The posterior probability distributions of ν_S, ν_R and ν_T

Figure 6.21 shows the posterior distributions for ν_S, ν_R and ν_T for the alignment of KAL153 HIV-1 DNA sequences described in Section 6.8.3. As the posterior distribution of ν_S peaks at values close to 1, it clearly shows that the model predicts only a few topology changes. ν_R was always relevant, where relevant in this context implies that the HIV-1 DNA sequence alignment was never modelled as a single state. This indicates that it is highly likely that the rate heterogeneity occurs – see Section 6.27. However, ν_R indicates more changes in the rate than expected, which may be due to problems related to the alignment of the HIV-1 DNA sequences, or some other mismatch between the assumptions of the model and what actually

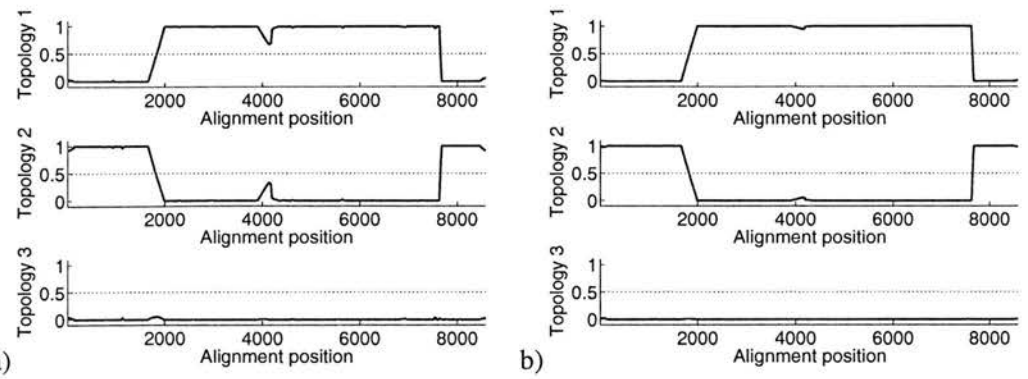


Figure 6.22: The posterior distribution of the phylogenetic tree topology along the alignment of HIV-1 DNA sequences, laid out in an identical fashion to Figure 6.11. Again, Sub-figure a) shows the predictions for the PRJ-FHMM while Sub-figure b) shows the predictions for the MCP of Minin et al. (2005). Using a letter to represent the sub-type, the topologies are Topology 1: [(F,A),(B,X)]; Topology 2: [(F,B),(A,X)] and Topology 3: [(F,X)(A,B)]. The predictions are very similar between the models, as both indicate that KAL153 (labelled X) is a recombinant of subtypes A and B, as found by Liitsola et al. (1998).

occurred in the evolution of these DNA sequences. Instead, we follow Minin et al. (2005) and set $C_R^{\min} = 0.999$ and $C_T^{\min} = 0.999$, a lower bound equivalent to the setting of Minin et al. (2005) where $\lambda_R^B = 9$. This is because their model is a special case of our model, with fixed values of ν_S and ν_R – see Equation (6.47). Since HIV-1 contains ten major genes along the alignment (*gag*, *pro*, *pol*, *env*, *vif*, *vpr*, *vpu*, *tat*, *rev* and *nef* – see Suchard et al., 2003), this corresponds to both the MCP and the PRJ-FHMM expecting on average each gene to have its own rate and transition-transversion ratio. After this constraining is done, the posterior probability of ν_T being relevant increases from 0.377 ± 0.01 to 0.766 ± 0.006. This indicates that it is likely that the transition-transversion ratio varies along the HIV-1 DNA sequence alignment.

6.9.3.2 Posterior distributions of the phylogenetic tree topology and rate

Figure 6.22 shows the predicted phylogenetic tree topology along the alignment. The predictions are very similar between the models, as both indicate that KAL153 is a recombinant of subtypes A and B, as found by Liitsola et al. (1998) and confirmed by Suchard et al. (2003). Suchard et al. (2003) finds another region of recombination occurring at around 4000-4200, which was not found by Minin et al. (2005). Interestingly, the PRJ-FHMM finds more support for this area of recombination than Minin et al. (2005). However, the probability of this region changing back to topology 2 is not significant enough to justify this conclusion.

Figure 6.23 shows the posterior distribution of the rate along the HIV-1 sequence alignment.

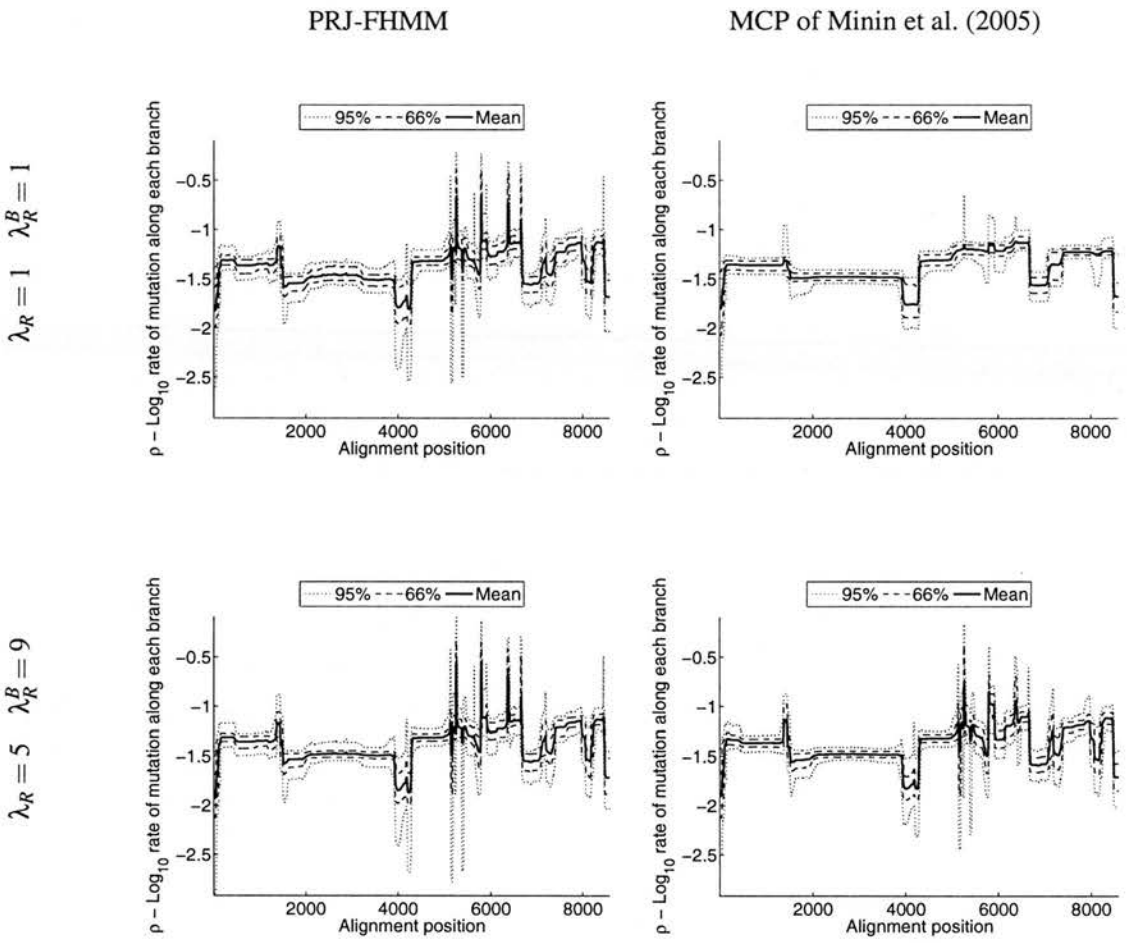


Figure 6.23: The credibility intervals of the posterior rate distribution along the alignment of HIV-1 DNA sequences, laid out in an identical fashion to Figure 6.12. When the MCP of Minin et al. (2005) and the PRJ-FHMM are given informative priors about the number of segments – each gene is expected to constitute a separate segment – they predict very similar patterns of rate variation.

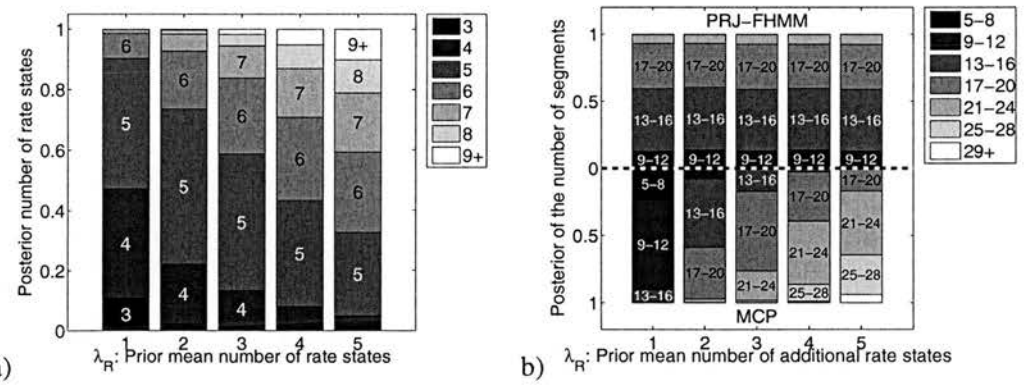


Figure 6.24: Comparisons of the predicted numbers of rate states and segments present on the alignment of HIV-1 DNA sequences by our PRJ-FHMM and the MCP of Minin et al. (2005), laid out in an identical fashion to Figure 6.13. The PRJ-FHMM model predicts that there are between 3 and 6 rate states.

While these predictions are difficult to verify, we show that when the MCP of Minin et al. (2005) and the PRJ-FHMM are given similar priors about the number of segments – each gene is expected to constitute a separate segment – they predict very similar patterns of rate variation. Changing the prior number of rate states in the PRJ-FHMM model appears to have little affect on the predicted rate along the alignment.

Figure 6.24 shows the posterior number of rate states found by the PRJ-FHMM model and the sensitivity of the methods to changes in the prior. For most settings of the prior, the PRJ-FHMM model predicts that there are between 3 and 6 rate states in the alignment. As the mean number of posterior rates is always larger than the setting of λ_R (the mean prior number of rate states), it provides a possible indication that our attempted priors might be too small, and that applying a second hyper-prior inference scheme might be useful here, in an equivalent fashion as is performed for the number of segments.

It is interesting to note that even when the PRJ-FHMM is allowed to predict as many segments as the MCP (as $0.999 < \nu_R$ is equivalent to $\lambda_R^B = 9$), it predicts a smaller number of segments. This can be seen by comparing the posterior number of segments of the MCP when $\lambda_R = 5$ (and thus $\lambda_R^B = 9$), to the predictions of the PRJ-FHMM.

6.9.3.3 Posterior distribution of the transition-transversion ratio.

Figure 6.25 shows the posterior transition-transversion predictions for both models. Both modes predict significant changes in the transition-transversion ratio along the alignment. The principal difference is that the MCP of Minin et al. (2005), probably due to its tying the estimation of the transition-transversion ratio to estimating the rate, predicts block like changes in the transition-transversion ratio.

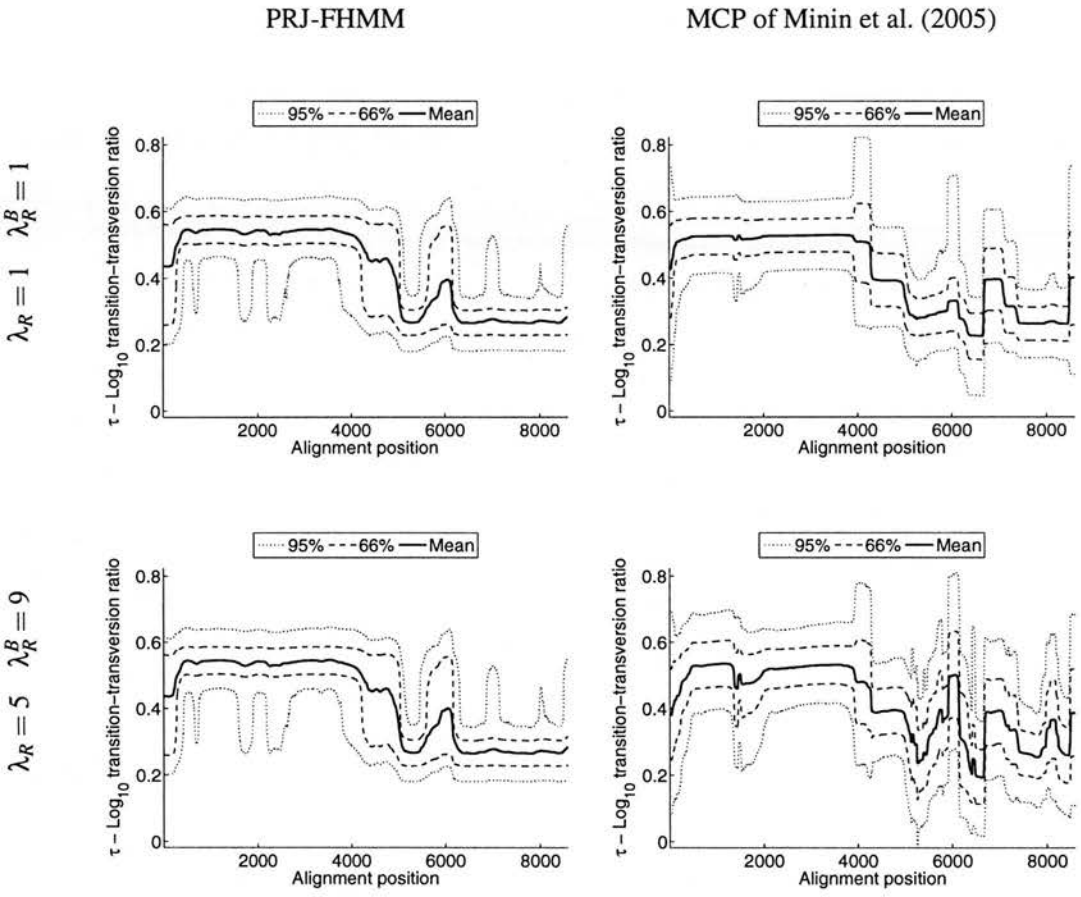


Figure 6.25: The credibility intervals of the posterior transition-transversion ratio along the alignment of HIV-1 DNA sequences, laid out in an identical fashion to Figure 6.14.

6.9.4 Investigating alternative priors on ρ_R and ρ_T

In Figure 6.26, we investigate the effect of changing the priors on ρ_R and ρ_T on the predicted rate along the alignment of Neisseria DNA sequences. The Gaussian prior gives a higher probability to the observed rates. As the set of observed rate states is more likely under the Gaussian prior, the PRJ-FHMM predicts more rate states, shifting the posterior distribution of the number of states to higher values. This may account for the slightly increased variance of the rate predictions.

Under the even-numbered order statistics prior, the ideal situations is that all states are spaced equally. Instead, in the rate case they cluster around -1.2, where this arragnement has a low prior probability. This results in ρ_R having the lowest expected number of rate states, and most conservative predictions for the rate. The uniform prior appears to be somewhere in the middle of the even-numbered order statistics prior and the Gaussian prior.

In Figure 6.27, we perform the equivalent investigation on the alignment of HIV-1 DNA sequences. The Gaussian prior removes the more extreme predictions as seen in the 95% confidence interval, as they are less supported by the prior. In contrast, the even ordered statistics prior encourages these extreme rate states, as can be seen by the more extreme spikes in the 95% credibility intervals. Again, the Gaussian prior gives the highest posterior number of rate states, while the even numbered order statistics gives the lowest posterior number of rate states.

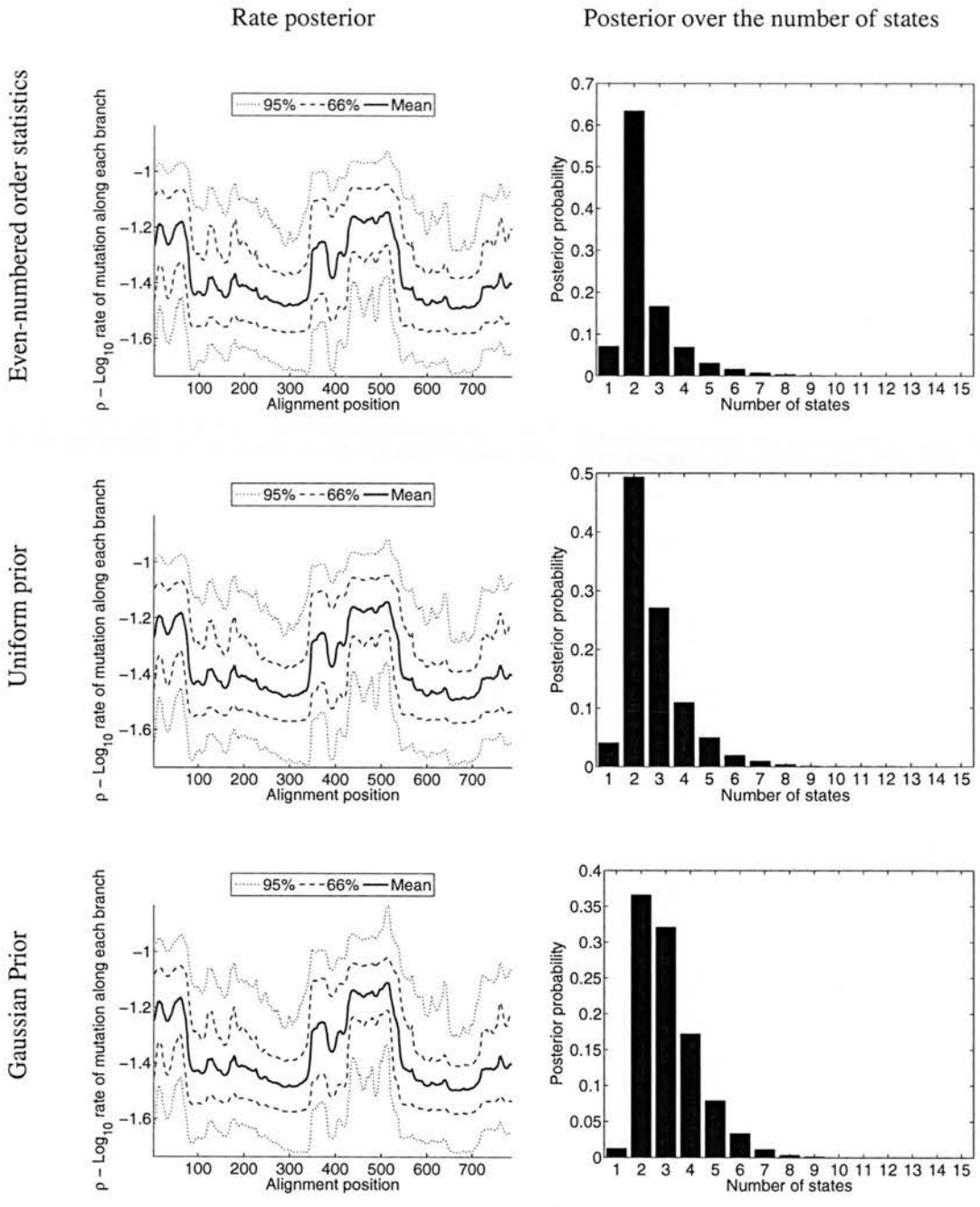


Figure 6.26: Investigating the effect of using different priors on $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ for the posterior distribution of the rate on the alignment of Neisseria DNA sequences. Panels on the left show the credibility intervals of the rate distributions, as seen in Figure 6.12, while the panels on the right show the posterior distribution of the number of states.

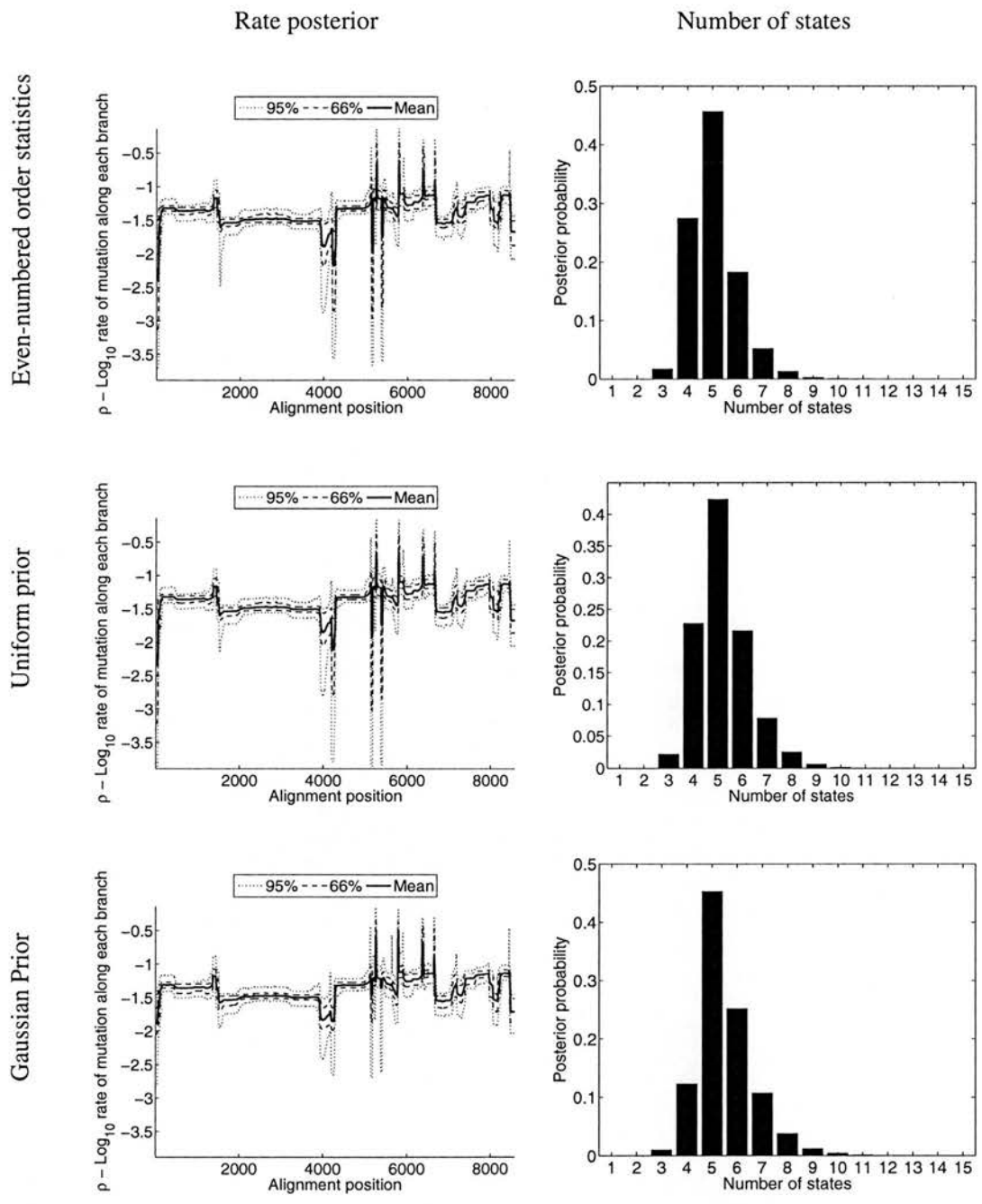


Figure 6.27: Investigating the effect of using different priors on $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ for the predictions on the alignment of HIV-1 DNA sequences, laid out in the same fashion as Figure 6.26.

6.10 Discussion

6.10.1 Why extreme rate states drive up the value of ν_R

In Husmeier (2005), the posterior distribution for ν_R was always found to peak at values close to 1, which indicates the presence of long rate segments. Here, we discuss how their model was limited to find such high ν_R values due to including an unrealistically high rate in the default set of rate states chosen. We offer two with two alternative explanations for this phenomenon.

For illustration purposes, assume that the emission probability associated with the unrealistically high predicted rate ρ^* is consistently equal to 0 and ignore the topology states $\mathbf{H}_S$ and transition-transversion ratio states $\mathbf{H}_T$. The computation of the marginal likelihood $P(\mathcal{D}|\nu_R)$ requires a marginalisation over all hidden state sequences:

$$P(\mathcal{D}|\nu_R) = \sum_{\mathbf{H}_R} \prod_t P(\mathbf{y}_t|H_{R,t}) P(H_{R,t}|H_{R,t-1}) . \tag{6.49}$$

Each hidden state sequence that transits into ρ^* has a zero contribution due to the fact that $P(\mathbf{y}_t|\rho^*) = 0$. Consequently, the marginal likelihood $P(\mathcal{D}|\nu_R)$ when including the extra hidden state is formally identical to the marginal likelihood without the extra hidden state except that the transition probabilities $P(H_{R,t}|H_{R,t-1})$ with $H_{R,t} \neq H_{R,t-1}$ are scaled down by a factor of $k/k+1$, where k is the number of different hidden states with the extreme rate state ρ^* included. Hidden state sequences with more transitions get more strongly penalised owing to the larger accumulation of penalising factors $k/k+1$. Larger marginal likelihood values can be achieved by giving a stronger weight to sequences with few state transitions, which is the case for high values of ν_R. This effect does not happen if an alternative non extreme rate state $\hat{\rho}$ is added, as transitions into this would not have such a large penalty term.

As an alternative explanation, consider that the marginal likelihood can be computed with Monte Carlo by sampling hidden state sequences from the prior distribution $P(\mathbf{H}_R|\nu_R)$ and weighting them by the likelihood $P(\mathcal{D}|\mathbf{H}_R)$ according to

$$P(\mathcal{D}|\nu_R) = \sum_{\mathbf{H}_R} P(\mathcal{D}|\mathbf{H}_R) P(\mathbf{H}_R|\nu_R) \tag{6.50}$$

This gives an estimator of the marginal likelihood as $P(\mathcal{D}|\nu_R) \approx \frac{1}{N} \sum_{i=1}^{N} P\left(\mathcal{D}|x^{(i)}\right)$ where $x^{(i)} \sim P(\cdot|\nu_R)$ is the i^{th} sample of the state sequence given ν_R. Low values of ν_R lead to more state transitions, and hence a higher probability of transiting into the extreme rate state ρ^*. However, upon transiting into ρ^*, $P(\mathcal{D}|\mathbf{H}_R)$ is zero. These sequences hence do not augment the accumulated Monte Carlo sum in Equation (6.50), while they augment the denominator and, hence, effectively incorporate a penalty effect. Consequently, low values of ν_R decrease the marginal likelihood and are effectively penalised against.

This behaviour was revealed when using the Reversible Jump as when the rate states are adjusted, these high rate states are lost, with the result that ν_R can deteriorate, allowing the

model to pick up on other short range behaviour like codon specific rate variation. This codon specific rate variation occurs because nucleotide sequences can code for protein sequences where successive triplets of nucleotides specify an amino acid – see Section 6.7.3 for more details of this codon effect. Modelling the transition matrix between states in more detail would remove this effect, which comes from the combination of only modelling a single transition probability ν_R between rate states and including unrealistic states.

6.10.2 Neisseria results

For the Neisseria DNA sequence alignment analysed in Section 6.9.1, the PRJ-FHMM has provided independent verification for the claim of Zhou and Spratt (1992) that the anomalous region around $507 - 538$ is not the result of recombination. Instead, the PRJ-FHMM and the MCP of Minin et al. (2005) have predicted that this anomalous region is part of a larger region of rate variation. In contrast to the MCP, we have also consistently identified rate variation occurring between alignment positions $1 - 75$, one of the more diverged regions found by Zhou and Spratt (1992). The MCP of Minin et al. (2005), on the other hand, did not consistently detect this region; the variability of its rate prediction results from the uncertainty in how to set λ_R^B. We have demonstrated that the proposed model infers the transition probability ν_R – equivalent to λ_R^B by Equation (6.47) – from the DNA sequence alignment, picking up long scale behaviour, and codon position specific rate variation, and we have shown that by setting the hyperparameters $\mathbf{C}$ appropriately, we can focus on the behaviour we wish to investigate. Additionally, the decoupling of segments and states allowed us to predict that there were two to three different types of rate states present along the alignment.

6.10.3 Maize results

In the analysis of the maize DNA sequence alignment in Section 6.9.2, we provided independent confirmation of the prediction of Husmeier and McGuire (2003) that gene conversion occurred in the maize actin genes. We additionally discovered a region which exhibits a region with significantly lower selective pressure (as seen in Figure 6.17). Husmeier and McGuire (2003) could not detect this region as their model cannot detect rate variation[2].

Furthermore, we have shown that this region, in addition to its high rate, does not exhibit a codon effect (see Figure 6.20). These findings indicate that it is possible that the original alignment inadvertently included an intron, as introns generally have a higher rate, and do not exhibit a codon effect. This may also explain the smaller magnitude of the codon effect in the maize alignment as opposed to the Neisseria alignment (compare Figures 6.10 and 6.15), as a part of the maize alignment does not exhibit a codon effect, reducing the support for this mode

[2]If examined closely, there is a very small blip in the topology predicted by Husmeier and McGuire (2003) where the rate variation occurs.

in the posterior distribution. This conjecture was independently confirmed in a discussion with a biologist (Frank Wright, personal communication).

6.10.4 HIV-1 results

For the HIV-1 DNA sequence alignment, we did not infer a clearly interpretable probability distribution of ν_R; this is in contrast to the distribution obtained for Neisseria, depicted in Figure 6.10b. A possible reason is inaccuracies in the DNA sequence alignment; owing to their large genetic diversification, HIV sequences are well-known to be intrinsically difficult to align. This points to an advantage of our proposed method over the MCP method of Minin et al. (2005), which does not include this inference step (λ_R^B, the parameter equivalent to ν_R, is set fixed) and hence lacks this diagnostic tool. However, when setting the value of ν_R to a fixed value corresponding to the value of λ_R^B used by Minin et al. (2005), our method effectively reproduces the authors' predictions.

The model of Husmeier (2005) is also unable to indicate a mismatch between the model and alignment due to the effect discussed in Section 6.10.1.

6.10.5 Comparison with the model of Husmeier (2005)

The extra flexibility gained from varying the rate states allows the PRJ-FHMM model to pick up on many different behaviours that occur in real sequence alignments. The model of Husmeier (2005) would never be able to infer solutions like the multi-modal posterior distribution for ν_R found for Neisseria in Figure 6.10. This is because their model, given a specific setting of the rates, appears to be limited to picking up on a single posterior ν_R mode. Furthermore, the model does not even provide indications that other solutions exist. Certainly, there is no way to set the rates such that all behaviours can be seen. This is made even more difficult by the non-obvious behaviour outlined in Section 6.10.1.

The extra flexibility has other advantages as well, like greater accuracy in characterising the rate along the alignment. This was demonstrated on a synthetic alignment in Figure 6.3. We show that using a single setting of the rates limits the accuracy of the predictions, and causes spurious predictions of rate variation.

Husmeier (2005) does not model the transition-transversion ratio as changing along a sequence. However, the transition-transversion ratio does vary along some alignments, like that of HIV-1, as shown in Figure 6.25 and already predicted by Minin et al. (2005).

The PRJ-FHMM model gives predictions of the number of states (or patterns or evolution) present, which the model of Husmeier (2005) is inherently incapable of. These patterns may lead to insight about what is occurring in the underlying biology, and allow a more detailed exploration of why this behaviour occurs. Discovering that multiple parts of the genome have

similar patterns of rate variation might reveal interesting hidden connections between these regions.

6.10.6 Comparison with the MCP of Minin et al. (2005)

Boys et al. (2000) outlined some of the advantages that HMMs have compared to breakpoint models. When, for instance, recombination occurs in the middle of a DNA sequence alignment, the PRJ-FHMM can easily identify that the segments on either side have identical characteristics by assigning them to the same state, and thus with every extra occurrence of the state increase the confidence of estimating the state characteristics. In contrast, the MCP independently estimates the rate and transition-transversion ratio for each repeated occurrence of a state because it is modelled as a separate segment. We investigated the effect of revisiting states on a synthetic alignment in Section 6.7.4, and found repeated state visitations can reduce the average error in the predicted rate, as seen in Figure 6.8. The PRJ-FHMM can find repeated occurrences of states in a computationally efficient manner due to the existence of efficient algorithms for inference in HMMs.

Additionally we have shown that the MCP of Minin et al. (2005) is effectively a special case of our model with a fixed value of ν_R, equivalent to the parameter λ_R^B, by Equation (6.47). In practice, there is uncertainty about how to set λ_R^B. The proposed Bayesian inference scheme addresses this uncertainty consistently by sampling λ_R^B from the posterior distribution. As seen from Figures 6.12 and 6.13, this leads to a considerable stabilisation of the predictions of the rate along the bacterial DNA alignment.

Fixing λ_R^B or ν_R has other disadvantages as well, as due to only using a single fixed value, Minin et al. (2005) cannot infer the multiple behaviours exhibited by the alignment of maize in Figure 6.15 and the alignment of Neisseria shown in Figure 6.10.

6.11 Conclusion

In this chapter, we proposed a fully Bayesian phylogenetic factorial hidden Markov model to simultaneously detecting recombination and characterise the rate states (or patterns of evolution) and rate segments of alignments of DNA sequences. This has many applications in functional genomics like identifying functional regions of proteins (Nimrod et al., 2005) and in comparative genomics (Chen and Blanchette, 2007) for detecting regulatory elements. In contrast, the focus of previous work such as Husmeier (2005) was on detecting recombination, and required choosing the set of possible rate states (or average branch lengths) in advance, leading to limited accuracy in characterising the rate, spurious predictions of rate variation, and an inability to analyse the rate states. We applied the model to a range of real and synthetic alignments to understand the properties of our proposed model, and contrast our proposed model to other

state of the art methods.

We have shown that the MCP model of Minin et al. (2005) is a special case of our HMM formulation, and that we have consistently addressed the uncertainty inherent in the breakpoint model about how to set the number of rate segments. This was shown on the bacterial DNA sequence alignment where the predictions of the MCP were highly dependent on the exact settings of the prior. In contrast, our predictions of the rate along the alignment were stable, and due to the decoupling of states and segments allowed us to predict that there are only two to three rate states present in agreement with the two anomalous regions found by Zhou and Spratt (1992).

Our model can be enhanced to exploit the annotations that are available for the alignments, namely that given the positions of introns and exons, the codon effect can be modelled by defining codon position specific rate offsets. This was suggested by Felsenstein and Churchill (1996b), but has not yet been integrated into our model and software.

There are many promising ways to exploit the model's ability to use considerably more complex state/segment specific evolutionary models, involving substantially more parameters than the HKY85 model employed in the present analysis. Consider using the model of Goldman and Yang (1994), which directly characterises the rate of nucleotide triplets with a detailed model involving 63 parameters; this will be almost unfeasible under a breakpoint model as each segment would require independently estimating this large number of parameters (see also Kosiol et al. (2007) for a recent 71 parameter model). In contrast our proposed model would use all repeated occurrences of a state to estimate these parameters, due to its HMM formulation.

6.12 Future Work

Other future work apart from that mentioned in the conclusion is to apply the model to very long alignments like to whole chromosomes. The more instances of each state, the better its properties can be estimated. This also increases the benefits of modelling the transition matrix in Equation (6.9) in more detail, as there will be more examples of transitions between the states.

The sharp rate peaks in the posterior distribution of the rate along HIV-1 seen in Figure 6.23 might be due to alignment issues. It may be possible to combine estimating mutation rates with attempting to discover flaws in the alignment, or to directly aligning the sequence while estimating the mutation rates. However, there are not yet any methods that simultaneously align a sequence, detecting recombination and rate variation, indicating that this is a difficult problem.

Symbol	Description
$\mathbf{y}_t$	Represents the t^{th}column in the alignment.
$\mathcal{D}$	The set of all columns from the alignment – $\mathcal{D} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$.
N	Number of columns in the alignment.
S	Represents the topology factor of the hidden Markov model.
R	As above, but represents the evolutionary rate factor.
T	As above, but represents the transition-transversion ratio.
$C_A^{\max}$	Maximum allowed value of ν_A.
$C_A^{\min}$	Minimum allowed value of ν_A.
$\mathbf{C}$	The set of all constraints - $\mathbf{C} = \left\{C_A^{\max}, C_A^{\min} \middle\| A \in \{S,R,T\}\right\}$.
ν_A	The probability than any two successive sites $H_{A,t-1}$ and $H_{A,t}$ are equal.
$\boldsymbol{\nu}$	The set of all ν_A variables – $\boldsymbol{\nu} = \{\nu_S, \nu_R, \nu_T\}$.
$\rho_{A,i}$	The i^{th} factor that is allowed in chain A.
$\boldsymbol{\rho}_A$	The list of different values that are explored in chain A, i.e. $\boldsymbol{\rho}_R$ is the set of list rate states.
k_A	The number of different states that the hidden chain A can assume, equal to the length of $\boldsymbol{\rho}_A$. For instance, k_R is the number of different rate states.
$H_{A,i}$	The hidden variable representing the factor A of the i^{th} column position.
$\mathbf{H}_A$	The chain of hidden variables of the factor A along the sequence – $\mathbf{H}_A = \{H_{A,1}, \ldots, H_{A,N}\}$.
$\mathbf{h}$	The set of all chains – $\mathbf{h} = \{\boldsymbol{\rho}_S, \boldsymbol{\rho}_R, \boldsymbol{\rho}_T\}$.
λ_A	The mean prior number of states in k_A, where $A \in \{R,T\}$.

Table 6.3: Notation used in this chapter. Unless stated otherwise, $A \in \{S,R,T\}$.

6.13 Chapter Conclusion

We have shown in this chapter that the we can with some degree of accuracy model the rate along a sequence alignment. This model is a substantial contribution to the field of phylogenetics in its own right. In the next chapter, we summarise the contributions of this thesis and proposed some future methods to exploit this extra information rate information for detecting the binding sites of the PRMs.

Conclusion

Chapter 7

Conclusion

7.1 Main contributions

In this thesis, we have proposed two novel alternative models for detecting the binding site motifs of Peptide Recognition Modules (PRMs), and a novel method for determining rate variation and recombination along sequence alignments. The principal contributions of this thesis are:

- Proposing (in Chapter 3) a novel discriminative model to distinguish between the closely related binding site motifs of different PRMs. We have also shown that this model outperforms an alternative, generative model.
- Proposing (again in Chapter 3) a novel hybrid model where the motifs from the generative model are further refined in a discriminative fashion. We further showed that this hybrid model achieved a better performance than either model on its own.
- Suggesting (in Section 3.9) many avenues of future research, ranging from model enhancements, methodological improvements, significant sources of alternative prior information about binding site, to possible applications of the model as an additional step to improve an alternative motif finding method.
- Proposing (in Chapter 7) a novel, computationally efficient discriminative model that incorporates information from all non binding sequences into discovering the binding motifs of PRMs.
- Proposing (in Chapter 6) a novel model for simultaneously detecting rate variation and recombination along DNA sequence alignments. Due to its RJMCMC inference scheme, this model can also infer the unknown set of rate and transition-transversion states.
- Illustrating on synthetic (in Section 6.7) and on real (see Sections 6.8 and 6.9) alignments of DNA sequences that state based models can be superior to breakpoint based models.

7.2 Future Work

We have already outlined many promising avenues of future research for enhancing the discriminative binding site models in Section 3.9 and for the phylogenetics model in Section 6.11. Ultimately, it is the combination of the methods proposed within the thesis that could provide the best performance. We will now outline some directions for future research.

7.2.1 Methods for enhancing motif searching by incorporating evolutionary context

Functionally important regions on proteins such as binding sites tend to be more conserved than the rest of the sequence (Nimrod et al., 2005). Hence, binding motifs are more likely to occur in conserved sequence regions, so we should bias our search towards such regions. The discriminative models described in Chapters 3 and 4 currently use the uniform binding prior asshown in Equations (3.6) and (4.3) respectively. These priors can be replaced by a prior such as:

$$P(a_{ij} = m) \propto \left(\prod_{k=1}^{p} 10^{-\langle H_{R,m+k-1} \rangle_{\cdot|\mathcal{D}}} \right)^{C}, \tag{7.1}$$

where $\langle f(X) \rangle_{\cdot|Y} = \langle f(X) \rangle_{X|Y} = \int_X P(X|Y) f(X) dX$ is the expectation operator, $\mathcal{D}$ is the alignment and $H_{R,i}$ is the $\log_{10}$ of the average branch length in i^{th} column along the alignment. $\langle H_{R,i} \rangle_{\cdot|\mathcal{D}} = \int H_{R,i} P(H_{R,i}|\mathcal{D}) dH_{R,i}$ is the posterior mean of the inferred $\log_{10}$ of the average branch length for the i^{th} column along the alignment. Hence, $10^{-\langle H_{R,i} \rangle_{\cdot|\mathcal{D}}}$ is the reciprocal of the branch length in i^{th} column along the alignment, and so more conserved regions end up with a higher prior. $H_{R,i}$ is inferred with the model proposed in Chapter 6. p is again the length of the motif, as in Chapters 3 and 4.

Equation (7.1) implicitly normalises over the different average conservation rates that might occur in alignments of sequences of different species – a region that is twice as converged as the rest of the sequence will still be twice as likely to contain the motif independent of the overall scaling of the branch lengths in the alignment of that sequence. Equation (7.1) also implicitly normalises over all putative motif positions.

The value assigned to C controls the effect of the rate conservation prior. Initially experimenting with $C = 1$ might be a good starting point, as this implies that regions that are twice as conserved are twice as likely to contain the binding site. Alternatively, C could be estimated by looking at how much more conserved some known motifs are compared with the generally observed, or suitable values for C could be determined using cross-validation.

This extra prior information should arguably increase the accuracy of the method by removing ambiguities about the true locations of the binding sites. It may also make the optimisation

process quicker and easier as the binding sites should be more obvious, leading to faster convergence.

The method of Reiss and Schwikowski (2004) can be updated to use this extra information. The prior on $a_{i,j}$ from their method can also be changed to Equation (7.1), thus favouring possible binding positions that are more conserved.

Alternatively, we can expand our input alphabet to incorporate information from the multiple sequence alignment. For instance, O'Rourke et al. (2005) use the information bottleneck method in an attempt to discretise the large amount of information available from a multi-alignment of an interesting sequence to a simple discrete alphabet. Their goal was allowing for sequence searches that are considerable quicker than using the full distribution. It would be a simple modification to use this method. However, they found that they lost information as compared to the full multiple sequence alignment. This also has the disadvantage of making it harder to apply the method to single sequences to locate the motif sites.

7.2.2 Incorporating structural information into the inference of rate variation

Nimrod et al. (2005) proposed an *in silicio* method to detect interaction surfaces on proteins. Their method requires the structure of a protein, and an alignment of that protein and a large collection of its homologues. These homologues are used to estimate a per position conservation score. The authors find these interactions surfaces through their conservation by "growing" a patch from a conserved site to any spatially close site that is also conserved. The aim to find the most conserved patches, as these are likely to correspond to functionally important areas like interaction surfaces. However, this is a heuristic method and an alternative is enhance the method that was proposed in Chapter 6. The links between the rate states of succesive positions alignment (as shown in Figure 6.1) model the fact that consecutive alignment positions are more likely to have the same same conservation rate. The model performance might be further improved by adding links between the rate states of peptides that are close in the protein structure but not consecutive. Then, the posterior probability of different alignment positions being in the same rate state would provide a rigorous method of identifying conserved patches.

Appendix A

The effect of the order constraint upon the prior

Jasra et al. (2005) claimed that imposing ordering constraints upon priors leads changes the resulting inference compared to performing the re-ordering after the simulation. Here, we show that this does not apply in our model. Consider two alternative prior distributions on the rate or transition-transversion ratios. Let A be either R or T. The ordered prior is:

$$P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_1) = \mathbb{I}(\rho_{A,1} \leq \rho_{A,2} \leq \cdots \leq \rho_{A,k_A})(k_A!)\prod_{i=1}^{k_A} Q_A(\rho_{A,i}), \tag{A.1}$$

where we introduce $\mathcal{H}_1$ to represent the assumption of an ordered prior. An alternative choice of prior distribution would be:

$$P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_2) = \prod_{i=1}^{k_A} Q_A(\rho_{A,i}), \tag{A.2}$$

where this choice of prior is represented by $\mathcal{H}_2$. We wish to show that the marginal posterior probability $P(\mathcal{D}|k_A)$ is unaffected by our choice of prior. Consider:

$$P(\mathcal{D}|k_A) = \int_{\boldsymbol{\rho}_A} P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_1)\,d\boldsymbol{\rho}_A, \tag{A.3}$$

where:

$$P(\mathcal{D}|\boldsymbol{\rho}_A,k_A) = \sum_{\mathbf{h}}\int_{\boldsymbol{\nu}}\int_{\boldsymbol{\rho}_{\{R,T\}\setminus A}} \sum_{k_{\{R,T\}\setminus A}} P\left(\mathcal{D},\mathbf{h},\boldsymbol{\nu},\boldsymbol{\rho}_{\{R,T\}\setminus A}|\boldsymbol{\rho}_A,\boldsymbol{\rho}_S,k_S\right) d\boldsymbol{\rho}_{\{R,T\}\setminus A}d\boldsymbol{\nu}, \tag{A.4}$$

where we drop the dependence on $\boldsymbol{\rho}_S$ and k_S to keep the notation simpler. All these summations and integrations are independent of our choice of prior on $\boldsymbol{\rho}_A$. We want to show that: $P(k_A|\mathcal{D},\mathcal{H}_1) = P(k_A|\mathcal{D},\mathcal{H}_2)$. This is equivalent to showing that $P(\mathcal{D}|k_A,\mathcal{H}_1) = P(\mathcal{D}|k_A,\mathcal{H}_2)$ as the priors on $\mathcal{D}$ and k_A do not depend on our choice of an ordered or unordered prior on $\boldsymbol{\rho}_A$. Consider that:

$$P(\mathcal{D}|k_A,\mathcal{H}_2) = \int\int\cdots\int P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_2)\,d\rho_{A,k_A}\cdots d\rho_{A,2}d\rho_{A,1} \tag{A.5}$$

We can split up this integral into various sub-integrals:

$$P(\mathcal{D}|k_A,\mathcal{H}_2) = \int_{\rho_{A,1}\leq\rho_{A,2}\leq\rho_{A,3}\leq\cdots\leq d\rho_{A,k_A}} \int_{\rho_{A,2}\leq\cdots\leq\rho_{A,k_A}} \cdots \int_{\rho_{A,k_A}} P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)\,P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_2)\, d\rho_{A,k_A}\cdots d\rho_{A,3}d\rho_{A,2}d\rho_{A,1} + \int_{\rho_{A,2}\leq\rho_{A,1}\leq\rho_{A,3}\leq\cdots\leq d\rho_{A,k_A}} \int_{\rho_{A,1}\leq\cdots\leq\rho_{A,k_A}} \cdots \int_{\rho_{A,k_A}} P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)\,P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_2)\, d\rho_{A,k_A}\cdots d\rho_{A,3}d\rho_{A,1}d\rho_{A,2} + \cdots . \tag{A.6}$$

There are $k_A!$ such order constrained integrals, each corresponding to a different possible permutation of $\boldsymbol{\rho}_A$. Each point in the unconstrained integral is in only one of the constrained integrals[1], and none of the constrained integrals include points that were not in the original integral. Hence, the sum of these constrained integrals is the same as the original unconstrained integral.

As we know that $P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)$ is invariant to the order of terms in $\boldsymbol{\rho}_A$, all these integrals are equal as they simply permutations of each other. Hence:

$$\begin{aligned} P(\mathcal{D}|k_A,\mathcal{H}_2) &= (k_A!)\int_{\rho_{A,1}\leq\rho_{A,2}\leq\rho_{A,3}\leq\cdots\leq d\rho_{A,k_A}} \int_{\rho_{A,2}\leq\cdots\leq\rho_{A,k_A}} \cdots \int_{\rho_{A,k_A}} P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)\,P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_2)\, d\rho_{A,k_A}\cdots d\rho_{A,3}d\rho_{A,2}d\rho_{A,1} = \\ &\int_{\rho_{A,1}\leq\rho_{A,2}\leq\rho_{A,3}\leq\cdots\leq d\rho_{A,k_A}} \int_{\rho_{A,2}\leq\cdots\leq\rho_{A,k_A}} \cdots \int_{\rho_{A,k_A}} P(\mathcal{D}|\boldsymbol{\rho}_A,k_A)\,P(\boldsymbol{\rho}_A|k_A,\mathcal{H}_1)\, d\rho_{A,k_A}\cdots d\rho_{A,3}d\rho_{A,2}d\rho_{A,1} \\ &= P(\mathcal{D}|k_A,\mathcal{H}_1). \end{aligned} \tag{A.7}$$

We have thus shown that the marginal likelihoods do not depend on the choice of ordered or unordered prior. It is also easy to show that the posterior distribution over the other variables in the system is not affected—— in a similar fashion.

[1] We will ignore that points where $\rho_{A,i} = \rho_{A,j}$ where $i \neq j$ occur in multiple integrals as this does not affect the result of the integration.

Glossary

Protein Proteins are polypeptides, i.e. consist of one or more chains of amino acids, and play an integral part in every cellular process. It is the arrangement of peptide chains that give the protein its shape, as different peptides have different properties. The interactions, and thus the functions of a protein are determined by its shape which is ultimately determined by the sequence(s) of peptides. Various post-translational process can affect their final shape, and thus function.

Amino acids The building blocks of a protein. Proteins are formed by condensation reactions between L-amino acids. There are in total twenty different amino acids, all of which share a common $H_2NCHRCOOH$ base, where R is the side chain that differs between the amino acids. These side chains have varying biochemical properties such as their charge, their polarity, their acidity, and whether they are hydrophobic or hydrophilic. This last property is often vital in determining the final structure of the protein.

Domain A domain is a structurally conserved subsequence that occurs on many different proteins, not necessarily consisting of consecutive amino acids, or amino acids only on one chain. Domains tend to be significantly longer than motifs.

DNA Deoxyribon Nucleic acid (DNA) is the principal means for long-term storage of information within in a cell. It consists of a long chains of nucleotides joined together. DNA is present in most cells as complementary double strands.

RNA RiboNucleic Acid (RNA) is similar to DNA, but contains ribose, instead of deoxyribose. RNA occurs only rarely as a double stranded, but single stranded RNA folds back on itself and can form many different structures. RNA is chemically more prone to hydrolysis than DNA, and thus is not generally used for long term storage of genetic information.

Nucleotide The building blocks of both DNA and RNA. They consist of a 5-membered ribose or deoxyribose ring linked to one of four bases. In DNA, the four bases are: thymine (T), cytosine (C), adenine (A) and guanine (G), while in RNA, thymine is replaced with uracil. Nucleosides are linked to the next by a phosphate group in a phosphodiester bond, where a nucleoside with a phosphate is called a nucleotide.

Ligand In general, any molecule that binds and forms a complex with another molecule. The other molecule is generally entitled the receptor. This binding generally changes the shape of the receptor, altering or enabling its function. Ligands and receptors play central roles in signalling pathways.

Receptor A protein that is bound by a ligand and then initiates a cellular process.

Sequence Motif Generally referred to simply as a motif, this is a short sequence of conserved nucleotides or amino acids that have some biological significance. They are usually detected by their statistical over-representation.

Alignment An alignment of sequences shows which nucleotides (or letters in general) correspond between the sequences. This is required when the sequences not only mutate, but also undergoes insertions and deletions. See Figure 5.2 for an example.

Topology A binary hierarchical structure describing the evolutionary relationships between a set of taxa (or species).

Recombination Recombination is process whereby different organisms swap or copy genetic information to each other. This can often be detected as a change in topology.

Mutation The process by which a nucleotide changes to another nucleotide over time. These mainly occur due to copying errors. Mutations can be point substitutions where a nucleotide changes to another nucleotide, insertions of extra nucleotides or deletions of existing nucleotides.

Mutation rate The average number of mutations observed between the sequences. Note that mutation is a random process, but in functionally important regions, mutations can severely impact the ability of the organism to pass on its genetic material. Hence, this will vary along alignment.

Rate variation Refers to changes in the observed mutation rate along sequences.

Codon Effect Each amino acid is coded for by a triplet of nucleotides. This gives a redundant code, and most of this redundancy is in the last nucleotide position. Hence, mutations here are less likely to affect the resulting protein and thus less likely to be selected against.

Bibliography

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (2001). *Molecular Biology of the Cell - Fourth Edition*. Garland Science.

Bailey, T. L. and C. Elkan (1995). The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol 3*, 21–29.

Baldi, P. and P. Brunak (1998). *Bioinformatics - The Machine Learning Approach*. Cambridge, MA: MIT Press.

Barash, Y., G. Elidan, N. Friedman, and T. Kaplan (2003). Modeling dependencies in protein-dna binding sites. In *RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology*, New York, NY, USA, pp. 28–37. ACM Press.

Barker, D. and M. Pagel (2005, June). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology 1*(1), e3–.

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference*. Ph. D. thesis, The Gatsby Computational Neuroscience Unit, University College London.

Ben-Hur, A. and W. S. Noble (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics 21*, i38–46.

Bishop, C. M. (2006, August). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Bock, J. R. and D. A. Gough (2001). Predicting protein-protein interactions from primary structure. *Bioinformatics 17*, 455–460.

Boni, M. F., D. Posada, and M. W. Feldman (2007). An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets. *Genetics 176*(2), 1035–1047.

Boys, R. J. and D. A. Henderson (2001). A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Computer Science and Statistics 33*, 35–49.

Boys, R. J. and D. A. Henderson (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics 60*, 573–588.

Boys, R. J., D. A. Henderson, and D. J. Wilkinson (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics 49*, 269–285.

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician 46*(3), 167–174.

Chen, H. and M. Blanchette (2007). Detecting non-coding selective pressure in coding regions. *BMC Evolutionary Biology 7*(Suppl 1), S9.

Cooper, G. M., E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res. 15*(7), 901–913.

Crandall, K. (1995). Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus. *J. Virol. 69*(4), 2351–2356.

Davis, F. P., H. Braberg, M.-Y. Shen, U. Pieper, A. Sali, and M. Madhusudhan (2006). Protein complex compositions predicted by structural similarity. *Nucl. Acids Res. 34*(10), 2943–2952.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Deng, M., S. Mehta, F. Sun, and T. Chen (2002). Inferring Domain-Domain Interactions From Protein-Protein Interactions. *Genome Res. 12*(10), 1540–1548.

Dohkan, S., A. Koike, and T. Takagi (2004). Prediction of protein-protein interactions using support vector machines. In *Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2004)*.

Down, T. A., C. M. Bergman, J. Su, and T. J. P. Hubbard (2007, January). Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Computational Biology 3*(1), e7–.

Down, T. A. and T. J. P. Hubbard (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucl. Acids Res. 33*(5), 1445–1453.

Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

Eden, E., D. Lipson, S. Yogev, and Z. Yakhini (2007, March). Discovering motifs in ranked lists of dna sequences. *PLoS Computational Biology 3*(3), e39–.

Enright, A. J., I. Iliopoulos, and N. C. O. C. A. Kyrpides (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature 402*, 86–90.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*, 368–376.

Felsenstein, J. (2001). The troubled growth of statistical phylogenetics. *Systems Biology 50*(4), 465–467.

Felsenstein, J. and G. A. Churchill (1996a). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*(1), 93–104.

Felsenstein, J. and G. A. Churchill (1996b). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*(1), 93–104.

Ferraro, E., A. Via, G. Ausiello, and M. Helmer-Citterich (2006). A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics 22*(19), 2333–2339.

Galtier, N. and M. Gouy (1998). Inferring patterns and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution 15*(7), 871–879.

Galtier, N., N. J. Tourasse, and M. Gouy (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science 283*, 220–221.

Gavin, A.-C., P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga (2006, March). Proteome survey reveals modularity of the yeast cell machinery. *Nature 440*(7084), 631–636.

Gelman, A. and D. B. Rubin (1992, November). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A.

Stanyon, J. Finley, R. L., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg (2003). A Protein Interaction Map of Drosophila melanogaster. *Science 302*(5651), 1727–1736.

Girolami, M. (2007). personal communication.

Goh, C., A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen (2000). Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology 299*, 283–293.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol 11*(5), 725–736.

Gomez, S. and A. Rzhetsky (2002). Towards the prediction of complete protein–protein interaction networks. *Pac Symp Biocomput. 7*, 413–424.

Gomez, S. M., W. S. Noble, and A. Rzhetsky (2003). Learning to predict protein-protein interactions from protein sequences. *Bioinformatics 19*(15), 1875–1881.

Grassly, N. C. and E. C. Holmes (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution 14*(3), 239–247.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Grimmett, G. R. and D. R. Stirzker (1994). *Probability and Random Processes*, Chapter 6. Oxford Science Publications.

Guimaraes, K., R. Jothi, E. Zotenko, and T. Przytycka (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biology 7*(11), R104.

Hasegawa, M., H. Kishino, and T. Yano (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*, 160–174.

Hayashida, M. and N. Ueda (2004). A simple method for inferring strengths of protein-protein interactions. *Genome Informatics 15*, 56–68.

Hayashida, M., N. Ueda, and T. Akutsu (2003). Inferring strengths of protein-protein interactions from experimental data using linear programming. *Bioinformatics 19*(90002), 58ii–65.

Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution 36*, 396–405.

Hesterberg, T., D. S. Moore, S. Monaghan, A. Clip-son, and R. Epstein (2005). *Introduction to the Practice of Statistics*, Chapter 14, pp. 14–14 – 14–18. W.H. Freeman & Company.

H.M.Berman, J.Westbrook, Z.Feng, G.Gilililand, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne (2000). The protein data bank. *Nucleic Acids Research 28*, 235–242.

Huang, C., F. Morcos, S. P. Kanaan, S. Wuchty, D. Z. Chen, and J. A. Izaguirre (2007). Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Husmeier, D. (2005). Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics 21*, ii166–ii172.

Husmeier, D. and K. Althoefer (1998). Modelling conditional probabilities with network committees: how overfitting can be useful. *Neural Network World 8*, 417–439.

Husmeier, D., R. Dybowski, and S. Roberts (2005). *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. New York: Springer.

Husmeier, D. and G. McGuire (2003). Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution 20*(3), 315–337.

Husmeier, D. and F. Wright (2001a). Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology 8*(4), 401–427.

Husmeier, D. and F. Wright (2001b). Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics 17*(Suppl.1), S123–S131.

Husmeier, D. and F. Wright (2002). A Bayesian approach to discriminate between alternative DNA sequence segmentations. *Bioinformatics 18*(2), 226–234.

Husmeier, D., F. Wright, and I. Milne (2005). Detecting interspecific recombination with a pruned probabilistic divergence measure. *Bioinformatics 21*(9), 1797–1806.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki (2001, April). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A 98*(8), 4569–4574.

Janin, J. (2002). Welcome to capri: A critical assessment of predicted interactions. *Proteins: Structure, Function, and Genetics 47*(3).

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling.

Jothi, R., M. G. Kann, and T. M. Przytycka (2005). Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics 21*, i241–250.

Keeler, J. D., D. E. Rumelhart, and W. K. Leow (1991). Integrated segmentation and recognition of hand-printed numerals. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky (Eds.), *Advances in Neural Information Processing Systems 3*, San Mateo, CA, pp. 557–563. Morgan Kaufmann Publishers.

Koike, A. and T. Takagi (2003). Prediction of protein-protein interaction sites and protein-protein interaction pairs using support vector machines. *Genome Informatics 14*, 500–501.

Kosiol, C., I. Holmes, and N. Goldman (2007). An Empirical Codon Model for Protein Sequence Evolution. *Mol Biol Evol.*

Krogan, N. J., G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt (2006, March). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature 440*(7084), 637–643.

Larget, B. and D. L. Simon (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution 16*(6), 750–759.

Lassmann, T. and E. Sonnhammer (2005). Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics 6*(1), 298.

Lattman, E. E. (2005). Sixth meeting on the critical assessment of techniques for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics 61*, 1–236.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science 262*, 204–218.

Lehrach, W. P., D. Husmeier, and C. K. I. Williams (2006a). A regularized discriminative model for the prediction of protein-peptide interactions. *Bioinformatics 22*(5), 532–540.

Lehrach, W. P., D. Husmeier, and C. K. I. Williams (2006b). Probabilistic in silico prediction of protein-peptide interactions. In E. Eskin, T. Ideker, B. Raphael, and C. Workman (Eds.),

Joint Annual RECOMB 2005 Satellite Workshops on Systems Biology and on Regulatory Genomics, Volume 4023 of *Lecture Notes in Computer Science*, pp. 188–197. Springer.

Leung, H. C. M. and F. Y. L. Chin (2005). Finding exact optimal motifs in matrix representation by partitioning. *Bioinformatics 21*, ii86–92.

Li, S. S.-C. (2005). Specificity and versatility of sh3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem. J. 390*(3), 641–653.

Liitsola, K., I. Tashkinova, T. Laukkanen, G. Korovina, T. Smolskaja, O. Momot, N. Mashkilleyson, S. Chaplinskas, H. Brummer-Korvenkontio, J. Vanhatalo, P. Leinikki, and M. O. Salminen (1998). HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS 12*, 1907–1919.

Liu, J. S., A. F. Neuwald, and C. E. Lawrence (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc. 90*(432), 1156–1170.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation 4*, 415–447.

MacKay, D. J. C. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation 11*(5), 1035–1068.

Margulies, E. H., M. Blanchette, D. Haussler, and E. D. Green (2003). Identification and Characterization of Multi-Species Conserved Sequences. *Genome Res. 13*(12), 2507–2518.

Margulies, E. H., G. M. Cooper, G. Asimenos, D. J. Thomas, C. N. Dewey, A. Siepel, E. Birney, D. Keefe, A. S. Schwartz, M. Hou, J. Taylor, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, J. B. Brown, P. Bickel, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, E. A. Stone, K. R. Rosenbloom, W. J. Kent, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, A. Hinrichs, H. Trumbower, H. Clawson, A. Zweig, R. M. Kuhn, G. Barber, R. Harte, D. Karolchik, M. A. Field, R. A. Moore, C. A. Matthewson, J. E. Schein, M. A. Marra, S. E. Antonarakis, S. Batzoglou, N. Goldman, R. Hardison, D. Haussler, W. Miller, L. Pachter, E. D. Green, and A. Sidow (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res. 17*(6), 760–774.

May, A. C. (1999). Towards more meaningful hierarchical classification of amino acid scoring matrices. *Protein Eng. 12*(9), 707–712.

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution 34*, 126–129.

Mayrose, I., N. Friedman, and T. Pupko (2005). A Gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics 21*(suppl_2), ii151–158.

McGuire, G. and F. Wright (2000). TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics 16*(2), 130–134.

McGuire, G., F. Wright, and M. Prentice (1997). A graphical method for detecting recombination in phylogenetic data sets. *Molecular Biology and Evolution 14*(11), 1125–1131.

McGuire, G., F. Wright, and M. Prentice (2000). A Bayesian method for detecting past recombination events in DNA multiple alignments. *Journal of Computational Biology 7*(1/2), 159–170.

Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics 21*(13), 3034–3042.

Moniz de Sa, M. and G. Drouin (1996). Phylogeny and substitution rates of angiosperm actin genes. *Molecular Biology and Evolution 13*, 1198–1212.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. A. Sigrist, R. Vaughan, and E. M. Zdobnov (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucl. Acids. Res. 31*(1), 315–318.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. A. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats (2007). New developments in the InterPro database. *Nucl. Acids Res. 35*(suppl_1), D224–228.

Nabieva, E., K. Jim, A. Agarwal, B. Chazelle, and M. Singh (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics 21*(suppl_1), i302–310.

Nabney, I. T. (2002). *NETLAB: algorithms for pattern recognition*. Springer-Verlag New York, Inc.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*, Volume 118 of *Lecture Notes in Statistics*. New York: Springer. ISBN 0-387-94724-8.

Neduva, V., R. Linding, I. Su-Angrand, A. Stark, F. d. Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell (2005, December). Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biology 3*(12), e405–.

Neduva, V. and R. B. Russell (2006, October). Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol 17*(5), 465–471.

Ng, P., N. Nagarajan, N. Jones, and U. Keich (2006). Apples to apples: improving the performance of motif finders and their significance analysis in the Twilight Zone. *Bioinformatics 22*(14), e393–401.

Ng, S.-K., Z. Zhang, and S.-H. Tan (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics 19*(8), 923–929.

Nimrod, G., F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko (2005). In silico identification of functional regions in proteins. *Bioinformatics 21*, i328–337.

O'Flanagan, R. A., G. Paillard, R. Lavery, and A. M. Sengupta (2005). Non-additivity in protein-DNA binding. *Bioinformatics 21*(10), 2254–2263.

O'Rourke, S., G. Chechik, R. Friedman, and E. Eskin (2005). Discrete profile alignment via constrained information bottleneck. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 1009–1016. Cambridge, MA: MIT Press.

Pavesi, G., G. Mauri, and G. Pesole (2001). An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics 17*(suppl_1), S207–214.

Pawson, T. and J. D. Scott (1997). Signaling Through Scaffold, Anchoring, and Adaptor Proteins. *Science 278*(5346), 2075–2080.

Pearl, J. (1988, September). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann.

Piero, F., P. Florencio, V. Alfonso, and R. Casadio (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem 269*(5), 1356–1356.

Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. pp. 61–74.

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile ; précédées des règles générales du calcul des probabilités.*

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez (2005). InterProScan: protein domains identifier. *Nucl. Acids Res. 33*(suppl_2), W116–120.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Ramani, A. K. and E. M. Marcotte (2003, 3). Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology 327*, 273–285.

Rambaut, A. and N. C. Grassly (1997). Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences 13*(3), 235–238.

Rasmussen, C. E. and C. K. I. Williams (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Reiss, D. J. and B. Schwikowski (2004). Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics 20*(suppl1), i274–282.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young (2000). Genome-Wide Location and Function of DNA Binding Proteins. *Science 290*(5500), 2306–2309.

Rosenberg, M. S., S. Subramanian, and S. Kumar (2003). Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol 20*(6), 988–993.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method.* New York: John Wiley & Sons.

Russell, R. B., F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali (2004). A structural perspective on protein-protein interactions. *Current Opinion in Structural Biology 14*, 312–324.

Sandve, G. and F. Drablos (2006). A survey of motif discovery methods in an integrated framework. *Biology Direct 1*(1), 11.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics 18*(3), 502–504.

Segal, E., Y. Barash, I. Simon, N. Friedman, and D. Koller (2002). From promoter sequence to expression: a probabilistic framework. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, New York, NY, USA, pp. 263–272. ACM Press.

Segal, E. and R. Sharan (2004). A discriminative model for identifying spatial cis-regulatory mdoules. In *RECOMB 2004 Conference Proceedings*, pp. 822–834.

Shoemaker, B. A. and A. R. Panchenko (2007, March). Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Computational Biology 3*.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res. 15*(8), 1034–1050.

Siepel, A. and D. Haussler (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology 11*(2-3), 413–428.

Skilling, J. (2006). Nested sampling for bayesian computations. In *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics*.

Sollich, P. and P. Krogh (1996). Learning with ensembles: How overfitting can be useful. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, Volume 8, pp. 190–196. The MIT Press.

Sprinzak, E. and H. Marglit (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology 311*, 681–692.

Stelzl, U. and E. E. Wanker (2006, December). The value of high quality protein-protein interaction networks for systems biology. *Current Opinion in Chemical Biology 10*(6), 551–558.

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer (2003). Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of the American Statistical Association 98*(462), 427–437.

Sudol, M. and T. Hunter (2000). New wrinkles for an old domain. *Cell 103*, 1001–1004.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences 17*, 57–86.

Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics 17*(12), 1113–1122.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res. 22*(22), 4673–4680.

Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu (2005, January). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotech 23*(1), 137–144.

Tong, A. H., B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone, and G. Cesareni (2002). A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules. *Science 295*(5553), 321–324.

Twyman, R. M. (2004). *Principles of Proteomics*. New York: BIOS Scientific Publishers.

Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg (2000, February). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature 403*(6770), 623–627.

von Mering, C., E. M. Zdobnov, S. Tsoka, F. D. Ciccarelli, J. B. Pereira-Leal, C. A. Ouzounis, and P. Bork (2003). Genome evolution reveals biochemical networks and functional modules. *PNAS 100*(26), 15428–15433.

Wang, H., E. Segal, A. Ben-Hur, D. Koller, and D. L. Brutlag (2004). Identifying protein-protein interaction sites on a genome-wide scale. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 1465–1472. Cambridge, MA: MIT Press.

Werhli, A., M. Grzegorczyk, M. Chiang, and D. Husmeier (2006). *Statistics in Genomics and Proteomics*, Chapter Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments, pp. 23–34. Centro Internacional de Matematica.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Comput. 7*(1), 117–143.

Yamanishi, Y., J.-P. Vert, and M. Kanehisa (2005). Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics 21*(suppl_1), i468–477.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol 10*(6), 1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol. 39*(3), 306–314.

Yang, Z. (1995). A Space-Time Process Model for the Evolution of DNA Sequences. *Genetics 139*(2), 993–1005.

Zhou, J. and B. G. Spratt (1992). Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology 6*, 2135–2146.